Data Science for the National Big Data R and D Initiative

Story

Slides

Slide 1 Data Science for the National Big Data R&D Initiative
Slide 2 Background
Slide 3 Data Science Knowledge Base in MindTouch
Slide 4 Data Science Data Publication: EPA Open Data Policy Inventory
Slide 5 EPA Public Excel File
Slide 6 EnviroAtlas: Link to Summary
Slide 7 Data Science Analytics in Spotfire
Slide 8 Some Results and Next Steps

Spotfire Dashboard
Research Notes

2015 NITRD Big Data Strategic Initiative Workshop

Home
Organization
Agenda

THURSDAY, JANUARY 22, 2015 - RECEPTION, RIGG'S LIBRARY
FRIDAY, JANUARY 23, 2015 - WORKSHOP, COPLEY HALL & HARARI

Registration
Hotel Info
FAQ
Contact

Data Science Seminar Series @ NSF

Big Data, Data Science, and other Buzzwords that Really Matter

Symposium on the Interagency Strategic Plan for Big Data: Focus on R&D

NITRD: National Big Data Strategic Plan-Summary of Request for Information Responses

Slide 1 Title Slide
Slide 2 Outline
Slide 3 National Strategic Plan RFI
Slide 4 Response Locations
Submissions received to the RFI on National Big Data R&D Initiative

AHA
ALCF
Amazon Web Services
Argonne data strategy
BNL
BNL Big Data Sharing
BNL-HEP-NP-Wenaus
BNL Integrated BigData Mining and Visual Analytics
BNL-NSLS2
CNM
Columbia
CSC
Dinov_UMich
EDC_Rescsanski
ESIP
ESnet
General Atomics
GTech
HDF
HP
Intertrust
IRIS
MITRE
NCDS
NERSC-SDF
Niemann
Request for Input (RFI)-National Big Data R&D Initiative

ACTION

SUMMARY

FOR FURTHER INFORMATION CONTACT

DATE

SUPPLEMENTARY INFORMATION

Description

Overview

Background

Objective

What We Are Looking For

Who Can Participate

Submission Format

The National Big Data R&D Initiative

Vision and Actions to be Taken

Vision Statement

GovLoop

About Us

Resources

Analytics

Big Data

WHY DATA IS YOUR MOST CRITICAL ASSET

THE FOUNDATION FOR DATA INNOVATION: THE ENTERPRISE DATA HUB

The Journey to Big Data & Analytics Adoption

Introduction

Chapter 1: Leveraging Your Most Critical Asset: Data

Chapter 2: Big Data’s Role in Countering Fraud

Chapter 3: Improving the Safety of Our Communities
Chapter 4: How Data Analysis Powers Smarter Care

Why Data is Your Most Critical Asset

Introduction

How IBM Can Help

Analytics 101: Defining Different Kinds of Analytics

Big Data Case Studies

Big Data for Medical Breakthroughs

How the Anderson Cancer Center applies big data to cancer research

Predicting Demand with Analytics

Lessons learned from Blizzard Ski’s ability to predict ski conditions with data

Improving the Environment

How the city of Dubuque, Iowa, reduced its carbon footprint

Keeping Communities Safe

How Miami-Dade County combated crime

Improved Health Services to Communities

How Alameda County has transformed health care with data

Fighting Tax Corruption with Data

How New York leverages big data to reduce tax fraud

Realizing the Promise of Big Data

10 Questions to Start Your Big Data Initiative

9 Steps to Starting Your Big Data Journey

1. Master the Art of the Possible
2. Don’t Reinvent the Wheel
3. Show the Business Value
4. Select a Mission-Critical Problem to Solve
5. Assess Capabilities against Requirements
6. Start Small and Have Clear Outcomes
7. Collaborate Across Government
8. Measure Impact and Develop ROI
9. Make Data Actionable: Tell Case Studies

Resources and More Reading

About GovLoop
Recognition
About IBM
Resources and More Reading

Coming Soon: Big Data’s Role in Countering Fraud

The Foundation For Data Innovation: The Enterprise Data Hub Report

Introduction
Figure 1: Powering the EDH – Apache Hadoop
Facilitating Converged Analytics & Data Governance
EDH In Action: Defense Intelligence Agency Case Study
Figure 2 Monitoring at the Tactical Operations Center
Figure 3 Searching Documents
How Can You Get Started
  Ask the Right Questions
  Start Small and Iterate
  Work with Experts
About Cloudera
About GovLoop
Access and Use of NASA and Other Federal Earth Science Data
Abstract/Agenda
Rant & Rave Title: Access and use of federal Earth science data
Notes from the panel at the summer meeting
  Earth Exploration Toolbook (EET)
  Experiences Accessing Federal Data
  Confessions of a data Hoarder
  Applications of satellite data used in the SatCam and WxSat mobile apps
Discussion
Attachments/Presentations
Details
  People
  Participants
Citation
Big Data in Materials Research and Development
Overview
  Authors
  Description
  Topics
  Publication Info
Front Matter
Overview
Workshop Themes
  DATA AVAILABILITY
  DATA SIZE: “BIG DATA” VS. DATA
  QUALITY AND VERACITY OF DATA AND MODELS
  DATA AND METADATA ONTOLOGY AND FORMATS
METADATA AND MODEL AVAILABILITY

CULTURE

Session 1: Introduction to Big Data

FRONTIERS IN MASSIVE DATA ANALYSIS AND THEIR IMPLEMENTATION

FIGURE 1 The elements of systematic analysis of massive data for NASA

IBM AND BIG DATA

BIG DATA FOR BIOSECURITY

DISCUSSION

Session 2: Big Data Issues in Materials Research and Development

PHYSICS IN BIG DATA

MATERIALS GENOME INITIATIVE AND BIG DATA

BOX 1 Materials Genome Initiative

FIGURE B-1 Conceptual representation of the MGI

BOX 2 Integrated Computational Materials Engineering

FIGURE B-2 ICME system that unifies materials information

GENERAL ELECTRIC EFFORTS IN MATERIALS DATA: DEVELOPMENT OF THE ICME-NET

FIGURE 2 ICME-Net for a disk forging use case.

SMART MANUFACTURING: ENTERPRISE RIGHT TIME, NETWORKED DATA, INFORMATION, AND ACTION

FIGURE 3 Smart manufacturing work flow.

DISCUSSION

Session 3: Big Data Issues in Manufacturing

DATA NEEDS TO SUPPORT ICME DEVELOPMENT IN DARPA OPEN MANUFACTURING

FIGURE 4 Critical elements of a rapid qualification system.

THE MATERIALS INFORMATION SYSTEM

FIGURE 5 Materials information system.

DISCUSSION

Session 4: The Way Ahead

LIGHTWEIGHT AND MODERN METALS MANUFACTURING INNOVATION INSTITUTE: IMPLICATIONS FOR MATERIALS, MANUFACTURING, AND DATA

FIGURE 6 Elements of the NNMI

DIRECTION OF POLICY

Footnotes

1
2
3
4
5
6
References
A Workshop Statement of Task
B Workshop Participants
C Workshop Agenda
  FEBRUARY 5, 2014
  FEBRUARY 6, 2014
D Acronyms

NEXT

---

**Story**

**Data Science for the National Big Data R&D Initiative**

I received an email that called attention to this important Federal Register Notice. I noticed that the link to the document was missing so I contacted the person for more information, Wendy Wigen, and she provided both the links to the [NITRD Web Page](https://semanticommunity.info/Data_Science/Data_Science_for_the_National_Big_Data_R_and_D_Initiative) and [PDF document](https://www.federalregister.gov/documents/2019/07/14/2019-14951/data-science-for-the-national-big-data-r-and-d-initiative). Incidentally, a Google search for the document also found our article entitled [Making the Most of Big Data](https://www.federalregister.gov/documents/2019/07/14/2019-14951/making-the-most-of-big-data) because we had followed the many NITRD Big Data activities in recent years.


The RFI Summary says:
This request encourages feedback from multiple big data stakeholders to inform the development of a framework, set of priorities, and ultimately a strategic plan for the National Big Data R&D Initiative. A number of areas of interest have been identified by agency representatives in the Networking and Information Technology Research and Development (NITRD) Big Data Senior Steering Group (BDSSG) as well as the many members of the big data R&D community that have participated in BDSSG events and workshops over the past several years. This RFI is a critical step in developing a cross-agency strategic plan that has broad community input and that can be referenced by all Federal agencies as they move forward over the next five to ten years with their own big data R&D programs, policies, partnerships, and activities. After the submission period ends and the feedback is analyzed, a workshop will be held to further discuss and develop the input received.

The guidelines says:

- It will not focus on individual agency plans and will not contain an implementation plan.
- We are not interested in specific research proposals or vendor offerings.
- Big data R&D policy recommendations are not relevant to this exercise.
- This RFI is open to all.
- Participants must also be willing to have their ideas posted for discussion on a public website and included with possible attribution in the plan.
- All responses must be no more than two (2) pages long (12 pt. font, 1" margins) and include the Submission Format below.

Our Submission is as follows:

- Who you are—name, credentials, organization: Dr. Brand Niemann, Director and Senior Data Scientist/Data Journalist, Semantic Community and Founder and Co-organizer of the Federal Big Data Working Group Meetup
- Your contact information: bniemann@cox.net
- Your experience working with big data and your role in the big data innovation ecosystem: Former Senior Enterprise Architect & Data Scientist with the US EPA, works as a data scientist, produces data science products, and publishes data stories for Semantic Community, AOL Government, & Data Science & Data Visualization DC Founded and co-organizes the Federal Big Data Working Group Meetup.
- Comments and suggestions based on reading the initial framework, The National Big Data R&D Initiative: Vision and Priority Actions, and guided by the following questions:
  ◦ What are the gaps that are not addressed in the Visions and Priority Actions document? A community of practice with a Mission Statement like the Federal Big Data Working Group has and that implements our version of the NSF Strategic Plan: Data Scientists, Data Infrastructure, and Data Publications using Semantic Web - Linked Open Data like the NITRD espouses.
  ◦ From an interagency perspective, what do you think are the most high impact ideas at the frontiers of big data research and development? The Data FAIRPort, NIH Data Commons, and Semantic Community Sandbox for Data Science Publications in Data Browsers.
  ◦ What new research, education, and/or infrastructure investments do you think will be game-changing for the big data innovation ecosystem? Data Mining all the research, education, and infrastructure investments for what has already been/is being paid for, especially for the agencies that have $100M or more in R&D funding per the so-called OSTP Holdren Memo, like the Federal Big Data Working Group Meetup is doing.
How can the federal government most effectively enable new partnerships, particularly those that cross sectors or domains? By participating in Communities of Practice on Data Science, Data Visualizations, Big Data, etc.

- A short explanation of why you feel your contribution/ideas should be included in the strategic plan: The Federal Big Data Working Group is what the Vision and Actions to be Taken say - a new partnership, a new gateway, and a long term sustainable activity.
- Examples, where appropriate: Federal Big Data Working Group Meetup and Below

As a further example, I am going to structure all this content (Web, PDF, email) so it is all data (Digital Government Strategy), mine this content for data subject matter experts and data sets (Open Data Policy), and prepare and visualize the data sets to answer the four questions a data scientist does when building data infrastructure, and data publications:

- How was the data collected?
- Where is the data stored?
- What are the results?
- Why should we believe them?

In fairness, there are other organizations that are attempting to fill the gap, like GovLoop, which sent me this recent email that prompted me to mine their web site for analytics, big data, and data science (not yet):

The Foundation for Data Innovation: The Enterprise Data Hub

To leverage data in its raw form to improve decision-making, organizations must overcome numerous obstacles. One of the major challenges is centrally managing data from disparate locations and in various file types.

The solution for many lies in an architecture called an enterprise data hub (EDH) which consolidates and stores data.

The Defense Intelligence Agency (DIA) has over 1,000 active data sources, manages 400+ apps and supports over 230,000 daily users with mission-critical needs. By adopting an EDH, the DIA is cutting down storage duplication, decreasing the time to deliver information and realizing cost savings. Learn more about the DIA’s Enterprise Data Hub in this report.

Download the New Report Now
Helping you lay the foundation for data innovation,
Team GovLoop

So I looked at that report (PDF) in detail and found the description, but not the answers to the four data science questions above. This seems to be a key differentiator between government (NASA/EPA example), academia (NAS example), and industry (GovLoop - Cloudera example) and the Federal Big Data Working Group Meetup!

Since the NASA/EPA example talks about EPA wanting to develop a big data strategy, I check the EPA Data Finder that Ethan McMahan did originally and found to my pleasant surprise the following:

This site provides access to EPA’s data sources organized by subject as well as links to EPA's Public Data JSON file, EPA's Public Excel file, and EPA's Environmental Dataset Gateway. Please tell us what EPA data you want
most we understand your data needs. You can visit EPA's Developer Central website to see how you can use EPA's data and services.

This site is part of EPA's coordinated Open Data efforts as described in White House Executive Order 13642 and the associated "Open Data Policy-Managing Information as an Asset" Memorandum Open Data Policy (M-13-13).

The EPA Public Excel File contains 3429 rows by 10 columns and those 10 columns contain some, but not all the "Common Core" Required Fields in the Open Data Policy, like Contact person’s name for the asset and email address. This file can be used in Spotfire for analytics to categorize and mine these data for drill-down into individual data sets and opportunities for federating multiple data sets.

This also illustrates how Project Open Data data discovery works by my documenting the click trail as follows:

• Start at: http://project-open-data.github.io/
• Scroll down to find Example Data Hubs: http://project-open-data.github.io/data-hubs/
• Scroll down to find Environmental Protection Agency: http://www.epa.gov/data/
• See EPA's Public Excel file and download it: http://www2.epa.gov/sites/production/files/2013-final.xlsx

The answers to the Data Science Publication Questions are:

• How was the data collected?
  ◦ We do not know yet?
• Where is the data stored?
  ◦ Excel Spreadsheet
• What are the results?
  ◦ Not Provided – We will do some analytics in Spotfire
• Why should we believe them?
  ◦ We do not know yet?

Some Results and Next Steps

• Semantic Community has provide a Response to the Request for Input (RFI)-National Big Data R&D Initiative citing the Federal Big Data Working Group Meetup as an example.
• We have built a knowledge base of examples: government (NASA/EPA Workshop), academia (NAS Workshop), and industry (GovLoop - Cloudera and IBM) to highlight a key differentiator, namely data science data publications that answer four key questions.
• We have used the Open Data Agency Data Hubs to find and use the EPA Public Excel File to illustrate this.
• The result is that EPA Public Excel File is insufficient to answer the four questions to produce a Data Science Data Publication, but the analytics have led us to the EPA EnviroAtlas “big data” sets (209) that can be used when the data have been reformatted for ease of use.

Some pertinent highlights from the NAS Big Data in Materials Research and Development workshop are:

• Daniel Crichton, Director, Center for Data Science and Technology, Jet Propulsion Laboratory, was also a member of the committee that wrote the NRC report Frontiers in Massive Data Analysis (2013), described some of the findings and recommendations of that study, and JPL’s excellent big scientific data program.
• Dr. Pitera described three representative projects in materials and big data:

https://semanticommunity.info/Data_Science/Data_Science_for_the_National_Big_Data_R_and_D_Initiative
Updated: Sun, 14 Jul 2019 10:27:14 GMT
Powered by mindtouch
Harvard’s Clean Energy Project. This is a distributed computing project that focuses on using big data for materials discovery. It seeks to identify new photovoltaic materials with higher efficiency rates. This involves large-scale calculations, along with data mining and analytics.

Pharmaceuticals. Pharmaceutical companies wish to mine patent data and medical literature to identify new relationships between drugs and disease.

Mining equipment. The goal of this project is to understand predictive equipment maintenance for heavy mining machinery.

Dave Shepherd, Program Manager, Homeland Security Advanced Research Projects Agency, Department of Homeland Security, explained that the goal of biosecurity is to alleviate any accidental or intentional release of pathogens or other causes of disease. He said that biosecurity is a predictive field, seeking to prevent disease exposure and spread by means of anticipation, and it requires a long-term perspective. Dr. McGrath pointed out that biosecurity is a large and difficult problem, larger in scope than what the materials community faces.

Michael Stebbins, Assistant Director for Biotechnology, Science Division, White House Office of Science and Technology Policy, pointed out that the biomedical research community has been approaching a crisis in that it has become virtually impossible to reproduce biomedical results. This is not because of widespread fraud, but because of a lack of access to protocols, underlying data, and process information. Dr. Stebbins pointed out that the only “currency” available today to a research scientist is the scientific publication in a journal. It may be possible to add a new currency: the actual data. Dr. Stebbins said that the initial pilot programs OSTP is considering will be free. The researchers (not the publishers) will deposit their data with a registered third party.

TO FOLLOW

Slides

Slide 1 Data Science for the National Big Data R&D Initiative

http://semanticommunity.info/

http://www.meetup.com/Federal-Big-Data-Working-Group/

http://semanticommunity.info/Data_Science/Federal_Big_Data_Working_Group_Meetup
Slide 2 Background

Background

- I received an email that called attention to this important Federal Register Notice and the Knowledge Base for this story evolved to include:
  - Request for Input (RFI)-National Big Data R&D Initiative
  - Email and Web
  - The National Big Data R&D Initiative
  - PDF
  - GovLoop (Web and PDF)
    - The Journey to Big Data & Analytics Adoption
    - The Foundation for Data Innovation: The Enterprise Data Hub Report
  - Access and Use of NASA and Other Federal Earth Science Data
  - Email and Web
  - Big Data in Materials Research and Development
    - Web and PDF

Slide 3 Data Science Knowledge Base in MindTouch

Data Science for the National Big Data R&D Initiative

Data Science Knowledge Base in MindTouch

Slide 4 Data Science Data Publication: EPA Open Data Policy Inventory

http://project-open-data.github.io/
http://project-open-data.github.io/data-hubs/
http://www.epa.gov/data/
http://www2.epa.gov/sites/production...013-final.xlsx

Data Science Data Publication:
EPA Open Data Policy Inventory

- Project Open Data to Find EPA Public Excel File
  - Start:
    - http://project-open-data.github.io/
  - Scroll down to find Example Data hubs:
    - http://www.epa.gov/data/portal/health
  - Scroll down to find Environmental Protection Agency:
    - http://www.epa.gov/data/
  - See EPA’s Public Excel File and download it:
- Answers to Data Science Publication Questions:
  - How was the data collected?
    - Yes, we reviewed and vetted.
  - Where is the data located?
    - Data from various sources.
  - What are the results?
    - No data.
  - Why should we believe them?
    - We do not know yet.
Slide 5 EPA Public Excel File


Slide 6 EnviroAtlas: Link to Summary

https://edg.epa.gov/metadata/catalog/search/resource/details.page?uuid=%7b0068AF7C-DE6C-4510-806A-BC35F333FE24%7d

Slide 7 Data Science Analytics in Spotfire

Web Player
Slide 8 Some Results and Next Steps

http://semanticommunity.info/Data_Science/Data_Science_for_EPA_EnviroAtlas

Some Results and Next Steps

- Semantic Community has provided a response to the request for input (RFI)-National Big Data R&D Initiative citing the Federal Big Data Working Group Meetup as an example.
- We have built a knowledge base of examples: government (NASA/EPA Workshop), academia (NAS Workshop), and industry (GovLoop - Cloudera and IBM) to highlight a key differentiator, namely data science data publications that answer four key questions.
- We have used the Open Data Agency Data Hubs to find and use the EPA Public Excel File to illustrate this.
- The result is that EPA Public Excel File is insufficient to answer the four questions to produce a Data Science Data Publication, but the analytics have led us to the EPA EnviroAtlas “big data” sets (209) that can be used when the data have been reformatted for ease of use.

Spotfire Dashboard

For Internet Explorer Users and Those Wanting Full Screen Display Use: Web Player Get Spotfire for iPad App

Media, iframe, embed and object tags are not supported inside of a PDF.

Research Notes

Here’s your copy of GovLoop’s newest guide: The Journey to Big Data and Analytics Adoption Chapter 1: Leveraging Your Most Critical Asset: Data.

Click Here to Download The Guide

Stay tuned — we’ve got more big data coming your way! By downloading Chapter 1 of The Journey to Big Data and Analytics Adoption Guide, you will automatically receive Chapters 2 - 4 as soon as they are available.

Ethan McMahon is back at EPA from a NASA detail and working on an EPA Big Data Strategy and trying to get some money!

Feel free to edit the ESIP page (http://commons.esipfed.org/node/2419) with your suggestions and please bring your experience and ideas to the discussion!

My Chat Posts:
to Ethan McMahon (privately):
Data.gov now uses Drupal with CKAN
See: https://docs.google.com/document/d/1...h.40kkiyxtpnau

to Ethan McMahon (privately):
I find that downloads are preferable to APIs

to Ethan McMahon (privately):
We have been working with Steven Kempler's ESIP ESDA Earth Science Analytics

to Ethan McMahon (privately):
We have a number of NASA, EPA, etc. Data Science examples that we could show you

to Ethan McMahon (privately):
We are doing a panel at the DGI Government Big Data Conference 2014 next week: http://www.digitalgovernment.com/ Age...ce--Exp.shtml

to Ethan McMahon (privately):
Data Science is analyzing, visualizing and storifying data sets to answer four questions: How was the data collected; Where is it stored; What are the results; and Why should we believe them?

to Ethan McMahon (privately):
Our Federal Big Data Working Group Meetup has an audience of over 350 members now: http://www.meetup.com/ Federal-Big-Data-Working-Group/

to Ethan McMahon (privately):
We could help you with your EPA Big Data Strategy because we have been analyzing "EPA Big Data". See Data Science for EPA EnviroAtlas.

---

2015 NITRD Big Data Strategic Initiative Workshop


January 23, Washington, DC

2015 Big Data Strategic Initiative Workshop --January 23 -- Georgetown University (in Red) Sponsored by the NITRD Big Data Senior Steering Group in conjunction with Georgetown University, this workshop will bring together a small number of academics and industry leaders across multiple disciplines to further inform the development of an effective Federal Big Data Research Agenda. Invitation only, see http://workshops.cs.georgetown.edu/BDSI-2015 for more information.

Home

Source: http://workshops.cs.georgetown.edu/B...2015/index.htm
The growth in scale, diversity, and complexity of data has increased the demand for understanding large amounts of heterogeneous data. This presents new challenges to the way data and information are used and managed. A need exists to understand ways to design systems for processing big data, develop sophisticated analytics using large-scale data, consider privacy issues arising from use of this massive data, and determine ways to teach big data analytics across disciplines.

OBJECTIVES

Given the large number of overlapping data challenges many federal agencies are facing, a National Big Data R&D Initiative was launched in March 2012. A steering group (the BDSSG) is drafting a framework and establishing a set of priorities for a Federal Big Data Research Agenda. One part of developing a strategic plan is to collect input from those in academia and industry that are working on big data research and new technologies. We propose having a one-day workshop to bring together a small number of academics and industry leaders across disciplines to further inform the development of an effective Federal Big Data Research Agenda. The invitees will cooperatively:

- Analyze and focus on the research and development issues of problems that are fundamental in making progress toward understanding complex, big data;
- Specify current and future areas where major breakthroughs appear possible;
- Identify needed collaborations (e.g., inter-disciplinary, academic-industry);
- Identify research initiatives and facilities needed to meet current and future challenges;
- Consider privacy challenges and directions that should be considered;
- Discuss educational directions, needs and challenges for integrating big data into the curriculum of different disciplines.

WORKSHOP DATES AND LOCATION

BDSI-2015 will begin with a dinner reception on Thursday, January 22nd at 6:00 PM. The workshop will take place on Friday, January 23rd from 9:00 AM - 5:00 PM. Workshop check-in/registration will take place from 8:30 AM - 9:00 AM on January 23rd. BDSI-2015 will take place at Georgetown University in Washington, DC.

THIS WORKSHOP IS SUPPORTED BY NITRD MEMBER AGENCIES THROUGH AN NSF GRANT. PI: LISA SINGH. ALL OPINIONS, FINDINGS, CONCLUSIONS AND RECOMMENDATIONS IN ANY MATERIAL RESULTING FROM
THIS WORKSHOP ARE THOSE OF THE WORKSHOP PARTICIPANTS AND DO NOT NECESSARILY REFLECT THE VIEWS OF THE NITRD AGENCIES.

Organization

Source: http://workshops.cs.georgetown.edu/B...5/steering.htm

ORGANIZERS

Lisa Singh, Associate Professor of Computer Science, Georgetown University
NITRD Big Data Senior Steering Group (BDSSG)

WORKSHOP STEERING COMMITTEE

David Logsdon, Senior Director for Public Sector, TechAmerica
Sarah Nusser, Vice President for Research, Iowa State University
Howard Wactlar, Vice Provost for Research Computing, Carnegie Mellon University

NITRD AGENCIES

Click here for list

Agenda

THURSDAY, JANUARY 22, 2015 - RECEPTION, RIGG’S LIBRARY

6:00 PM - 6:30 PM    Networking

6:30 PM - 6:45 PM    Welcome
Lisa Singh, Georgetown University
Bob Groves, Georgetown University
Chaitan Baru, NSF

6:45 PM - 7:00 PM    Keynote
Tom Kalil, White House Office of Science & Technology Policy

7:00 PM - 7:45 PM    Table Discussions

7:00 PM - 7:45 PM    Closing Remarks
Allen Dearry, NIH
FRIDAY, JANUARY 23, 2015 - WORKSHOP, COPLEY HALL & HARARI

8:30 AM – 9:00 AM  Registration; continental breakfast

9:00 AM - 9:30 AM  Welcome & Workshop Goals

9:30 AM - 10:15 AM  Keynote
Andrew Moore, Carnegie Mellon University

10:15 AM - 10:30 AM  Break

10:30 AM - 11:30 AM  Panel: Future Directions for Big Data Research, Development & Education

Magdalena Balazinska, University of Washington
Data Science Education & Big Data Management

Michael Carey, University of California, Irvine
Big Data Systems

Kate Crawford, Microsoft Research and MIT Center for Civic Media
Data & Ethics

Sastry G. Pantula, Oregon State University
Statistics & Big Data

11:30 AM - 12:00 AM  Future Directions Discussion

12:00 PM - 1:00 PM  Lunch (provided by workshop)

1:15 PM - 3:15 PM  Breakout Sessions

3:15 PM - 3:30 PM  Break

3:30 PM - 4:30 PM  Plenary break out group reports

4:30 PM - 5:00 PM  Final Wrapup

Registration

Source: http://workshops.cs.georgetown.edu/registration.htm

MY NOTE: By Invitation Only
Hotel Info


MY NOTE: Not Applicable

FAQ


MY NOTE: Not Applicable

Contact


Wendy Wigen (NITRD)
wigen@nitrd.gov

Lisa Singh (Georgetown University)
singh@cs.georgetown.edu

Data Science Seminar Series @ NSF

Big Data, Data Science, and other Buzzwords that Really Matter

Source: Email

Wednesday, January 21, from 11:30-12:30
National Science Foundation, 4201 Wilson Blvd, Arlington, VA
Room 110, Stafford I (Public access is available, no security required)
Webinar: Registration Link

Michael Franklin, UC Berkeley Computer Science

Michael Franklin from Berkeley will be the inaugural speaker at the Data Science Seminar Series @ NSF organized by the AAAS Science and Technology Policy Fellows. This series is a monthly one-hour informational presentation and discussion about Data Science, Big Data and the Internet of Things as addressed by government agencies, academia, industry, and non-profit sectors. These informative presentations are open to all and meant to be timely and adaptable to current Data Science activities and the needs of our community.

Wednesday, January 21, from 11:30-12:30
National Science Foundation, 4201 Wilson Blvd, Arlington, VA
Room 110, Stafford I (Public access is available, no security required)
Webinar: Registration Link

Abstract

Data is all the rage across industry and across campuses. While it may be tempting to dismiss the buzz as just another spin of the hype cycle, there are substantial shifts and realignments underway that are fundamentally changing how Computer Science, Statistics and virtually all subject areas will be taught, researched, and perceived as disciplines. In this talk I will give my personal perspectives on this new landscape based on experiences organizing a large, industry-engaged academic Computer Science research project (the AMPLab), in helping to establish a campus-wide Data Science research initiative (the Berkeley Institute for Data Science), and my participation on a campus task force charged with mapping out Data Science Education for all undergraduates at Berkeley.

Biography

Michael Franklin is the Thomas M. Siebel Professor of Computer Science and Chair of the Computer Science Division of the EECS Department at UC Berkeley. He is the director of the Berkeley AMPLab, a 70+ person effort fusing scalable computing, machine learning, and human computation to make sense of data at scale. AMPLab software including: Spark, Shark, and Mesos, plays a significant role in the emerging Big Data ecosystem. The lab is funded by an NSF CISE Expeditions Award, the Darpa XData program, and 26 companies including founding sponsors Amazon Web Services, Google, and SAP.

Please feel free to distribute and post.

For additional information, please contact Nishal Mohan at nmohan@nsf.gov

Symposium on the Interagency Strategic Plan for Big Data: Focus on R&D

Source: Email October 7, 2014 (Word)

Eleventh Meeting of the Board on Research Data and Information

National Research Council

National Academy of Sciences Keck Building, Room 100

500 Fifth Street, NW

Washington, DC 20001

October 23, 2014

The Big Data Senior Steering Group (BDSSG) was initially formed to identify big data R&D activities across the Federal Government, offer opportunities for agency coordination, and jointly develop strategies for a national initiative. The National Big Data R&D Initiative (“the Big Data Initiative”) was launched in March 2012. Since the launch, the BDSSG has held numerous meetings and workshops, including a major showcase event of dozens of partnerships that will help
advance the frontiers of big data R&D across the country. Many participating federal agencies have already established new big data programs, policies and activities and plan to do more in the future. Currently, the BDSSG is drafting a framework and establishing a set of priory goals for a National Big Data R&D Strategic Plan.

The BDSSG is currently gathering information from multiple sectors to inform the development of an effective National Big Data R&D Strategic Plan. After the submission period ends on November 14 and the feedback is analyzed, the BDSSG will hold a workshop to further discuss and develop the input received. Additional detailed information about this Request for Information may be found at: https://www.nitrd.gov/bigdata/rfi/02102014.aspx.

This BRDI Symposium will be held in two sessions, beginning at 8:45. The first one will have presentations by six federal government speakers, beginning with an overview and followed by more focused talks in the five strategic areas of the Big Data Initiative: technologies; knowledge to action; sustainability; education; and gateways. The second session will feature three panelists from academia and industry, who will comment on the interagency Big Data Initiative and respond to the presentations in the first session. There will be ample time for discussion with the audience as well. The Symposium will be moderated by the co-chairs of the Board on Research Data and Information, Clifford Lynch of the Coalition for Networked Information, and Francine Berman of the Rensselaer Polytechnic Institute.

The Symposium will be followed immediately by the presentations of the award winners of the BRDI Data and Information Challenge. There is no fee to attend, but please contact the Board director, Paul Uhlir at puhlir@nas.edu or at 1 202 334 1531, to register in advance. Additional information about both these events, including the detailed program, is available at http://www.nas.edu/brdi.

Open Session

8:45 Session chair: Clifford Lynch, CNI (all session speakers not confirmed)

Overview Presentation: Allen Dearry, NIH
1. Technologies: Howard Wacltar, NSF*
2. Knowledge to action: Peter Lyster, NIH
3. Sustainability: Sky Bristol, USGS*
4. Education: Michelle Dunn, NIH
5. Gateways: Kamie Roberts, NIST

10:15 Break

10:30 Symposium (continued)
Session chair: Francine Berman, RPI
Panel discussion (all session panelists not confirmed)
Kelvin Droegemeier, University of Oklahoma*
Daniel Goroff, Alfred P. Sloan Foundation*
Elizabeth Grossman, Microsoft*

11:30 The BRDI Data and Information Challenge Awards

In Memoriam: Tribute to Lee Dirks Elizabeth Grossman, Microsoft
Presentations by the Winners
First Place: Adam Asare, Immune Tolerance Network, ITN Trial Share: Enabling True Clinical Trial Transparency
Runner up: Mahadev Satyanarayanan, Carnegie Mellon Univ., Olive: Sustaining Executable Content Over Decades

Presentation of Awards
12:25 End of meeting

* Not yet confirmed

NITRD: National Big Data Strategic Plan-Summary of Request for Information Responses

Source: https://www.nitrd.gov/bigdata/rfi/11...aStratPlan.pdf

Response Coordination Team
NSF AAAS Fellows: Renata Rawlings-Goss, Alejandro Suarez, Martin Wiener, Lida Beninson
NSF Staff: Fen Zhao, Dana Hunter
NITRD Staff: Wendy Wigen

Slides

Slide 1 Title Slide
Slide 2 Outline

- Introduction: Demographics
- Summary of Responses
  - Next generation Capabilities
  - Data to Knowledge to Action
  - Access to Big Data Resources
  - Education and Training
  - New Partnerships
  - Data Sustainability
  - Gateways
- Additions and Comments
  - Gaps
  - Game Changers
  - Security and Privacy

Slide 3 National Strategic Plan RFI

National Strategic Plan RFI

- We received 38 responses
- From 27 stakeholders including:
  - Industry
  - Academia
  - National Labs
  - Nonprofits
  - Community Leaders
  - Consortium
Slide 4 Response Locations

Response Locations

Slide 5 Next Generation Capabilities

Next Generation Capabilities

Vision 1: Create next generation capabilities by leveraging emerging Big Data foundations, technologies, processes, and policies. (62%)

- **Data Efficiency**
  - "Algorithm Acceleration with state-of-the-art hardware"
- **Infrastructure/Hardware**
  - "Infrastructure R&D should address delivery of information through varied, multiple, and concurrent channels and mobile devices"
- **Testbeds**
  - "Create cloud-based test beds for R&D with sample datasets, community supported algorithm libraries, with cross-agency access and support."
- **Data co-design**
  - "...design process where scientific problem requirements influence architecture design and technology, and constraints inform formulation and design of algorithms and software."

*Multiple responses*
Slide 6 Data to Knowledge to Action

Data to Knowledge to Action

Vision 2: In addition to supporting the R&D necessary to create knowledge from data, emphasize support of R&D to understand trustworthiness of data and resulting knowledge, and to make better decisions and breakthrough discoveries and take confident action based on them. (46%)

- Time to Knowledge
  - “...all processing could be done “inline,” with the data stream rather than as a post processing step”

- Data Integrity/Stewardship/Curation
  - “Develop a digital “chain-of-handling” to allow all relying parties to understand how, when, and by whom the results were derived. This should include code and parameters to aid in scientific reproducibility and post facto validation”
  - “…if data are not adequately curated, it is questionable whether they should have been generated in the first place”

Slide 7 Access to Big Data Resources

Access to Big Data Resources

Vision 3: Build and expand access to the Big Data resources and cyberinfrastructure - both domain specific and shared - that are needed for agencies to best achieve their mission goals and for the country to innovate and benefit. (51%)
Slide 8 Education and Training

Education and Training

Vision 4: To educate the next generation of scientists and engineers and fulfill increasing demand for analytical talent and capacity for the broader workforce (27%)

- Develop PhD fellowships for students who excel in either the foundations or exploitations of data science; allow longer time to joining faculty or industry to learn both technology and data in different disciplines
- Encourage public programs and certifications for roles at data-driven organizations, government competitions and prizes
- Leverage online learning
- Create hubs to enable curricula development
- Create graduate fellowships to add analytical skills
- Provide domain relevant data/software for teaching materials
- Redesign high school and collegiate science curriculum to incorporate data analytics

Slide 9 New Partnerships

New Partnerships

Vision 5: Foster the creation of new partnerships that cross sectors and domains (43%)

- Encourage inter-laboratory and cross-divisional collaboration
- Fuse government and social media data to draw intelligence
- Work with Nat’l Research Council to conduct unbiased assessment on strategic scientific data investments
- Partner with independent consulting firms to handle data policy compliance
- Devise industry incentives to develop solutions
- Create hubs to foster public-private collaborations
- Invest in the open source community
- Commercial partnerships for foundational software
Slide 10 Data Sustainability

Data Sustainability

- Long-term investments required for all aspects of cyber-infrastructure
  - New/definite funding models needed
- Stability and reliability through transitions
- Integration with library and info science communities
- Formation of new cross-agency entity
- Consideration of costs of archival of all data
- Consolidate data lifecycle management across agency resources

Slide 11 Gateways

Gateways

- Portals for sharing large data sets and analysis tools
- Scalable, API-driven software services
- Diversity of software packages and stacks
- Built with end-user in mind
  - Hide underlying infrastructure
  - Accessible with little/no programming experience
- Implementation in conjunction with data centers
Slide 12 Gaps

Gaps

What are the gaps that are not addressed in the Vision and Priority Actions Document? (79%)

- Organization/Mobility
- Data Bottleneck
- Long-term sustainability
- Energy of storage, processing and transmission.
- Metadata – Argonne National Lab
- Pace of analysis tool development
- Economic value
- Privacy
- Data Migration

Slide 13 Game-changers

Game-changers

What new research, education and/or infrastructure investments do you think will be game-changing for the big data innovation ecosystem? (21%)

- Use of J/TB (Joule per Terabyte) as a metric
- Cost of data storage
- Disaster Forecasting, Health sensors/records, education tracking
- Fusion of government and social media
- Big Data Research Centers
- Incentivization
- Big Data Infrastructure (on par with supercomputing centers)
Slide 14 Security, Privacy & Data Governance

Slide 15 Response Coordination Team

Submissions received to the RFI on National Big Data R&D Initiative

Source: [https://www.nitrd.gov/bigdata/rfi/02102014.aspx](https://www.nitrd.gov/bigdata/rfi/02102014.aspx)

My Note: See full text versions of ODF files below for analysis

- AHA
- ALCF
- Amazon Web Services
- Argonne data strategy

https://semanticommunity.info/Data_Science/Data_Science_for_the_National_Big_Data_R_and_D_Initiative

Updated: Sun, 14 Jul 2019 10:27:14 GMT

Powered by mindtouch
The Big Data SSG would like to thank all responders to this RFI. The submissions will be used to inform the development of an effective National Big Data R&D Strategic Plan.
AHA

Big Data is one of the biggest challenges in this century. The benefits of leveraging new sources of data have created many opportunities and serve as the impetus to usher in the next era of computing. I commend the BDSSG for assuming leadership and the arduous task of guiding our nation to tackle this exciting new problem. Identifying a strategic direction is critical in directing research efforts, which will form the fallow ground for new technologies to be born. Additionally, I am appreciative for the opportunity to provide feedback to this strategic plan and hope that the ideas I present to you in this letter create value in your effort.

Before beginning my feedback, I would like introduce my company and myself. My name is Juan Deaton, PhD. I’m a research scientist for AHA Products Group, which has created data compression and coding products for over two decades. AHA has noticed the many new challenges with storage, processing, and transmitting Big Data. One of our major goals has always been to create new hardware that will increase performance and energy efficiency of complex algorithms.

As you may know, we have reached a critical point in large scale computing. Given the trends outlined in 2009, large scale computing is consuming approximately 3% of the US total power, more power than the entire paper industry. In fact, power is the single largest expense in any data center, attributing to over 30% of operating costs. Big Data presents us with critical energy challenges that have real environmental consequences. Given the explosion of information generated by electronic and social media, today’s technologies simply cannot keep up with the rate that data is increasing. About 80% of the world’s data was created in last four years and by 2020 there is a predicted 40x increase in data production. For these reasons, I would urge the BDSSG to include investments in foundational research and cyberinfrastructure for higher performing and energy efficient to address storage, processing, and transmission of data.

We’ve found, that an increase in performance and energy efficiency is often achieved simultaneously when approached with the correct technology. This especially true for data compression. With data compression, systems can can increase throughput and storage capacity by representing the information with fewer bits. However, while CPUs have formed the basis for modern computing and software, they are not as suitable for all computing tasks. In fact, many data centers do not use compression because CPUs can consume more energy and have slower performance when performing compression. However, when data compression is implemented onto specialized hardware, performance and energy efficiency accelerates by a factor of 10x over CPUs. When used in a data center, compression accelerators can increase storage and throughput by factors of three or more. With accelerator hardware, applied to applications such as compression, data centers can realize critical performance increases with little additional energy cost. We believe there are other opportunities in big data applications where accelerators can increase efficiency and performance of these systems.

While specialized hardware will enable short-term performance increases, new approaches to storage, processing, and transmission of data are also necessary as well. I’ve outlined some future technologies below in addition to my short-term recommendations.

Short Term Recommendations. Support data centers research evaluating and developing accelerators, such as data compression or encryption, which simultaneously increase efficiency and performance. Support data center research on storage, transmission, and/or computational efficiency using JÆB (Joule per Terabyte) as a metric. Support top 10 most energy efficient data center lists/competition with recognition and financial motivation. Support research for...
developing hardware accelerators for big data applications and networked hardware accelerators for tera-scale operations.

Long Term Recommendations: Support research that will create new energy efficient technologies that can store, process, and transmit exabytes of data using the same amount of energy today's technologies. Support research in large-scale data compression, encryption, and erasure coding techniques that operate on terabytes of data.

Thank you for your time. It is my hope that these recommendations will create research to insure our energy security, a clean environment, and advancements in technology for the betterment of mankind.

Juan D. Deaton, PhD


ALCF

Michael E. Papka, ALCF Director -- papka@anl.gov


Submitted on behalf of the Argonne Leadership Computing Facility Argonne National Laboratory, Argonne, IL 60439

The Argonne Leadership Computing Facility (ALCF) is one of two DOE Leadership Computing Facility (LCF) centers in the nation dedicated to open science. Supported by the DOE’s Advanced Scientific Computing Research program, the LCF operates centers at Argonne National Laboratory and Oak Ridge National Laboratory, and deploys two diverse high-performance computer architectures that are 10 to 100 times more powerful than systems typically available for open scientific research. The LCF provides world-class computational capabilities to the scientific and engineering community to advance fundamental discovery and understanding in a broad range of disciplines. ALCF supports over 1,100 users and roughly 250 projects. ALCF users have generated more than 5 petabytes (PB) of data this year and ALCF manages over 13 PB of user data on disk and 14 PB on tape. For more information about the ALCF, visit alcf.anl.gov.

What are the gaps that are not addressed in the Visions and Priority Actions document?

Hierarchical and Distributed Data Management. A comprehensive management system to optimize data movement and placement between the various levels and locations of the storage hierarchy would both accelerate data-intensive applications and reduce or eliminate the need for programmers to develop custom data management schemes. Development of such a system will require the cooperation of storage vendors, operating systems developers, and applications programmers.
Co-Managing Computing and Long-Term Storage. Scientific datasets are growing ever--larger, and moving them around for simulation or analysis has become increasingly impractical. Instead, middleware and algorithms should manage the optimal placement of data in order to carry out the calculations requested. Research and development of new co--management techniques where compute, network, and storage are dynamically managed to provide the overall optimization, including predictive analysis of overall systems behavior, is essential. Dedicated facilities for the long--term storage of carefully curated data should also be established.

Trusted Data. Investment is required to ensure end--to--end data integrity, validation, recoverability, access control, and security are part of the Big Data ecosystem. Expecting each application development team to independently integrate these technologies is unrealistic. Instead, these capabilities should be provided by underlying software libraries and integrated into the hardware infrastructure. Likewise, significant work is needed to develop end--to--end security that does not inhibit scientific data analysis or sharing, but still provides the required level of security and access control.

From an interagency perspective, what do you think are the most high impact ideas at the frontiers of big data research and development?

Metadata. Metadata describes the content and structure of a particular dataset, and is critical to making that data available beyond the scientific group that created it. As experimental, sensor, and simulation datasets become larger, more complex, and potentially useful to a broader set of scientists, rich and well--structured metadata is essential. Domain scientists must be encouraged to invest time and effort into making their datasets easier for others to use, which vastly increases the value of the data and the potential for interagency collaboration.

Portals. Because the movement of large datasets is expensive and time consuming, co--locating accessible compute capability with data storage increases the potential of their scientific usefulness. However, most potential data users are not programmers skilled in the use of supercomputers or complex datasets. Web--based data portals are the best way to make these datasets widely usable. Computing facilities like the ALCF must invest in the infrastructure and support needed to host and integrate these interfaces.

What new research, education, and/or infrastructure investments do you think will be game--changing for the big data innovation ecosystem?

Investment in Data Facilities. DOE National User Facilities generate massive experimental, observational, and simulation datasets that are critical to increasing scientific understanding in a variety of fields. Investments at facilities like the ALCF are required to make this data widely available through structured data management and the availability of experts to help in data retrieval and analysis.

How can the federal government most effectively enable new partnerships, particularly those that cross sectors or domains?

Data User Facilities. There is the need for federally funded national user facilities for data that parallel the current national facilities for computation. These data user facilities would be dedicated to the curation, storage, and dissemination of important data sets. Such facilities must have the computing capabilities to provide analysis and visualization without requiring excessive data movement. In addition, these facilities will ensure the integrity and security
of the data, and make it available to a broad range of scientific and industrial projects. The federal government should also establish clear rules for data ownership and requirements for data sharing.

**Summary.** Facilities like ALCF are at the forefront of applying advanced computing, storage, and networking technology to achieve breakthroughs in science and engineering. By enhancing the ability of facilities like ALCF to provide open access to important datasets, along with the computing time and the human expertise needed to make use of these datasets, the federal government will usher in a new era of scientific discovery that leverages existing data in novel ways.

---

**Amazon Web Services**

Amazon Web Services, Inc. 12900 Worldgate Dr. Suite 800 Herndon, VA 20170 Jamie Baker Sr. Civilian Account Manager, AWS Public Sector 301-461-0569 bakjames@amazon.com Cage Code: 66EB1 DUNS Number: 965048981 NAICS: 518210

This Amazon Web Services, Inc. (AWS) package is provided for informational purposes only. The services included in this package are standard commercial services and due to the unique nature of the services, AWS’s standard terms and conditions must govern any use of and access to AWS services and are available at http://aws.amazon.com/agreement/. This package may include a set of suggested solutions for this opportunity that are based on our limited information, and should not be construed as a binding offer from AWS. For current prices for AWS services, please refer to the AWS website at http://www.aws.amazon.com.

In addition, the information contained herein may include technical data, the export of which is restricted by the U.S. Arms Export Control Act (AECA), as amended (Title 22, U.S.C. Sec 2751, et seq.), Export Administration Act of 1979, as amended (Title 50, U.S.C. App.2401.et seq.) or similar laws.

The purpose of this document is to provide the Networking and Information Technology Research and Development (NITRD) Big Data Senior Steering Group (BDSSG) with feedback to inform the development of a framework, set of priorities, and ultimately a strategic plan for the National Big Data R&D Initiative. We will restrict our feedback on the initial framework document, The National Big Data R&D Initiative: Vision and Priority Actions, to a few major items.

1. **Who are you, contact information, and what is your experience working with big data, and your role in the big data innovation ecosystem?** – Amazon Web Services, Inc. (AWS) is a leading commercial cloud provider in the USA, and a leading platform for development and operations of open source and commercial Big Data solutions. More information here: [http://aws.amazon.com/big-data/](http://aws.amazon.com/big-data/). Contact information can be found on this document’s cover page. In addition to a robust set of foundational infrastructure services, we provide robust security capabilities and managed services used widely in the commercial and academic research Big Data space.

2. **What are the gaps that are not addressed in the Visions and Priority Actions document?** – The initial framework document touches on the building and expansion of access to cyberinfrastructure, but does not explicitly relate this point to a later goal of long-term sustainability of the program, and resources created within such programs. AWS strongly believes that an operational expenditure (OpEx) approach to IT infrastructure increases the scalability, security and flexibility of resources available to end users, while also adding the benefits of multi-site collaborations through cloud technology.
The benefits of cloud computing include:

- Self-service provisioning;
- On-demand control and elasticity;
- No up-front capital expenses or commitments for infrastructure/facilities;
- Pay-as-you-go pricing;
- Improved agility; and,
- Lower cost of ownership than traditional on-premises offerings.

3. What do you think are the most high impact ideas at the frontiers of big data research and development?

What new research, education, and/or infrastructure investments do you think will be game-changing for the big data innovation ecosystem? – Big Data systems have the following characteristics: developed at a fast pace; infrastructure is based on commodity hardware and refreshed often; processes are tuned for horizontal scale. Commercial clouds, like AWS, offer an infrastructure platform that is well suited to all of these characteristics by providing rapid access to flexible and low cost IT resources in an on-demand, pay-as-yougo model with no upfront investments in hardware. Instead, you can provision exactly the right type and size of computing resources needed to solve the problem at hand. With Cloud Computing, you can access as many resources as you need, almost instantly, and only pay for what you use. The following are five advantages and benefits of adopting cloud computing:

- Trade capital expense for variable expense – Instead of having to invest heavily in data centers and servers before you know how you’re going to use them, you can use cloud computing and only pay when you consume computing resources, and only pay for how much you consume.
- Benefit from massive economies of scale – By using cloud computing, you can achieve a lower variable cost than you can get on your own. Because usage from hundreds of thousands of customers are aggregated in the cloud,
cloud computing providers such as AWS can achieve higher economies of scale which translates into lower pay as you go prices.

- **Flexibility – Stop guessing capacity** – Eliminate guessing on your infrastructure capacity needs. When you make a capacity decision prior to deploying an application, you often either end up sitting on expensive idle resources or dealing with limited capacity. With Cloud Computing, these problems go away. You can access as much or as little as you need, and scale up and down as required with only a few minutes notice.

- **Increase speed and agility** – In a cloud computing environment, new IT resources are only a click away, which means you reduce the time it takes to make those resources available to your developers from weeks to just minutes. This results in a dramatic increase in agility for the organization, since the cost and time it takes to experiment and develop is significantly lower.

- **Stop spending money on building, running, and maintaining data centers** – Focus on projects that are core to your organization, not the infrastructure. Cloud computing lets you focus on your own customers, rather than on the heavy lifting of racking, stacking and powering servers.

4. **How can the federal government most effectively enable new partnerships, particularly those that cross sectors or domains?** – Federal agencies can promote the creation of central data sharing hubs among the members of a particular initiative on commercial cloud infrastructures. This model would enable collaborators to bring their own software, processes, and financial resources to the same environment as the data hub. In this model, AWS’s cloud infrastructure would host high-value data sets, preconfigured software, and reference architectures that enable researchers within a community to bootstrap their research.

5. **A short explanation of why you feel your contribution/ideas should be included in the strategic plan.** – As the leading commercial cloud provider in the big data ecosystem, we feel that we have a depth of experience and knowledge to offer to the NITRD computing infrastructure at the scale of Big Data, while controlling for costs of the data analysis. We would welcome the opportunity to provide the NITRD BDSSG detailed information on considering an OpEx approach to IT infrastructure, the scalability, security and flexibility of the cloud, and the benefits of researchers collaborating through cloud technology.

Use or disclosure of data contained on this sheet is subject to the restriction on the title page of this response.

### Argonne data strategy

**Enabling Big Data in Science**

Mihai Anitescu anitescu@mcs.anl.gov Charlie Catlett catlett@anl.gov Ian Foster foster@anl.gov Salman Habib habib@anl.gov Mark Hereld hereld@anl.gov Paul Hovland hovland@mcs.anl.gov Jed Brown jdbrown@mcs.anl.gov Kate Keahey keahey@mcs.anl.gov Nicola Ferrier nferrier@anl.gov Michael Papka papka@anl.gov Robert Ross rross@mcs.anl.gov Rajeev Thakur thakur@mcs.anl.gov Tom Peterka tpeterka@mcs.anl.gov Venkat Vishwanath venkatv@mcs.anl.gov Stefan Wild wild@anl.gov Mike Wilde wilde@mcs.anl.gov Justin Wozniak wozniak@mcs.anl.gov

Technology advances have led to tremendous increases in the amount of science data generated from experimental facilities, observational platforms, and computational simulation.

The techniques used to manage, understand, and share these data have not kept pace. As a result, the scientific discovery process is being impeded, with engineers at facilities cobbling together systems to manage these data,
scientists wrestling with antiquated tools to analyze data, and collaborators struggling to share data and knowledge with one another.

A successful solution to the challenges of big data in science must encompass four key areas: • State of the art algorithms • High productivity environments • Robust software infrastructure • Capable facilities

Four components of a successful data intensive science strategy.

![Four components of a successful data intensive science strategy.](https://semanticommunity.info/Data_Science/Data_Science_for_the_National_Big_Data_R_and_D_Initiative)

**Algorithms**

With the explosion of data generated by experiments, observational platforms, and simulations, new techniques are required to efficiently reduce data, to manage data with uncertainty and/or with variable quality, to integrate scientific knowledge into analysis, and to operate at very large scale, on a variety of architectures. Our goal is to develop novel models and algorithms for analyzing science data in highly parallel and resource-constrained environments, keeping “science in the loop”. This approach will bring together math advances, CS techniques, and domain knowledge to remove scalability barriers of classical algorithms, produce the highest quality results, improve interactivity for prototyping, and enable automation of complex analysis.

Examples of ongoing activities include the activity by S. Wang et al. in the analysis of fluorescence microscopy data from the APS and work by T. Peterka et al. in the analysis of cosmology simulation data for the purpose of studying gravitational lensing and morphology of cosmic structure. Both of these activities highlight our approach of engaging local domain science staff in early stages of R&D to validate approaches while solving real problems.

**Environments**
Data science involves complex, resource-intensive, end-to-end processes that may span computers, buildings, facilities, and countries. Systems enabling design and execution of the complex workflows that describe the data science process are beyond state of the art in most cases. Our goal is to provide an end-to-end environment for data intensive science that provides effective views of scientists’ data, facilitates automation of common tasks, and executes these tasks efficiently on large scale data resources.

One example of activity in this area is our enhancement and deployment of the Galaxy Platform for use in genome sequencing, proteomics, and Earth science applications. Through these deployments we have gained increasing understanding of the needs of small teams of scientists, pointing the way towards general tools for larger communities. Another aspect of this problem is in orchestrating complex scientific workflows, and our Swift parallel scripting language is creating new capabilities for executing complex workflows on a variety of platforms, including leadership computing systems.

**Software Infrastructure**

Software infrastructure is critical to the effective management of data in many science projects, yet there is a huge disparity between science projects in terms of their infrastructure, its maturity, and its capabilities. The ultimate success of data intensive science in the DOE rests on our collective ability to create reusable infrastructure that support numerous, diverse teams. Our goal is to develop a shared, large-scale, heterogeneous data science infrastructure that is capable of supporting many science domains at the multi-institution level.

Our activities in the KBase project inform us as to some of the requirements of multi-institution software infrastructure, for example highlighting the need for scheduling of computation based on data locality as well as the capabilities of specific sites.

**Facilities**

Finally, rapidly increasing data volumes and subsequent complexity of analysis are already creating situations where science teams simply do not have the hardware to unlock the knowledge buried in their data. Leading-edge data intensive science facilities will provide the data analysis and storage capabilities, as well as the connectivity to instruments and other sites, that can enable science communities to pursue discovery via data intensive science.

In conclusion, our objective is to foster the growth of a vigorous data science community. We see these four areas as necessary and complementary components of a “data science substrate” that will accelerate discovery in the sciences by allowing researchers to work with data in terms of familiar abstractions, through convenient interfaces, and with state of the art analysis techniques.

---

**BNL**

**Accelerating Datadriven Innovation and Scientific Discovery**

Robert J. Harrison, rharrison@bnl.gov, (631) 37447090 Brookhaven National Laboratory, P.O. Box 5000, Upton, NY 119735000

---

https://semanticommunity.info/Data_Science/Data_Science_for_the_National_Big_Data_R_and_D_Initiative

Updated: Sun, 14 Jul 2019 10:27:14 GMT

Powered by mindtouch™
Brookhaven National Laboratory (BNL) is a multipurpose research institution funded primarily by the U.S. Department of Energy’s Office of Science. Located on the center of Long Island, New York, BNL brings worldclass facilities and expertise to the most exciting and important questions in basic and applied science—from the birth of our universe to the sustainable energy technology of tomorrow. We operate largescale facilities for studies in physics, chemistry, biology, applied science, and a wide range of advanced technologies. The success and benefit to the nation of these facilities and associated research programs are predicated upon our continued innovation of new techniques and technologies to manage, mine, and analyze data at the frontier of scientific discovery.

For instance, datacentric computation at BNL is central to discoveries at the CERN Large Hadron Collider (LHC) that is the largest scientific enterprise worldwide in data intensive computing. BNL’s ATLAS Tier 1 Center (the largest outside CERN) delivers costeffective computing to many physics programs, and BNL also provides the PanDA (Production and Distributed Analysis) system that orchestrates ATLAS’ data intensive computing at our computing facility and around the world: about 1.3 Exabytes were processed by PanDA in 2013. The ATLAS data set emerging from the first LHC run is currently about 150 PB. BNL physicists at our Relativistic Heavy Ion Collider (RHIC), the only collider now operating in the U.S., leads the world in exploring how the matter that makes up atomic nuclei behaved just after the Big Bang. These and multiple other capabilities make BNL one of the largest Big Data / High Throughput Computing resources and expertise pools in US science.

Data is the lifeblood of all our science programs, not just highenergy and nuclear physics. BNL is involved in all main DOE cosmology experiments. There are two unique features of cosmological datasets: first the information of interest is buried in correlations (e.g., power spectra and bispectra) within the petascale data sets, and second, cosmological datasets offer a great opportunity for serendipitous discovery but this is hard to automate and the data volume vastly exceeds what can be done even with citizen scientists – there are not enough humans around. In Climate, Environment, & Biosciences, BNL seeks to understand the relationships between climate change, sustainable energy initiatives, and the planet’s natural ecosystems. For instance, we host the Atmospheric Radiation Measurement (ARM) Climate Research Facility that operates and hosts the data from strategically located in situ and remote sensing observatories to improve the understanding and representation of clouds and aerosols as well as their interactions and coupling with the Earth’s surface. BNL plays a lead role in the DOE Knowledge Base (KBase) that is a bigdata bioinformatics collaboration developing tools for storing, interrogating and analyzing the exponentially increasing body of genome and metagenome sequences. A new frontier is emerging at the brand new National Synchrotron Light Source II (NSLSII) that uses electrons accelerated around a ring at nearly the speed of light to create beams of light in the xray, ultraviolet, and infrared wavelengths. It will generate many thousands of times more data than its predecessor being 10,000 times brighter with much larger/faster detectors and new instruments. Analysis of the generated petascale data will pave the way to discoveries in physics, chemistry, and biology — advances that will ultimately enhance national security and help drive the development of abundant, safe, and clean energy technologies. These are just a few of our interests that also include energy security, smart grids and energy infrastructure, remote sensing, cyber security, urban science, and so on.

In what must government invest to ensure progress towards broad science objectives while fully leveraging the huge economic engine of big data in business and industry that will continue to push technology in most of this space?

- Interagency research programs directed towards the intellectual foundations of translating science and engineering data into knowledge. Data is primarily a deep intellectual challenge, not just technology. For instance, new math and algorithms are necessary to efficiently and robustly identify correlations in peta/exascale data sets, to solve the inverse problems underlying advanced experiments, to robustly fuse data from multiple sources, or to identify novel features representing unexpected discoveries.
Industry-academic-government partnerships are needed to fill in the gap between what is needed and what is feasible with offtheshelf technologies. Motivating this are the investments currently underway by the Department of Energy in pursuit of exascale high-performance computing. The equivalent program is needed for big scientific data.

The continuum between big data and big compute needs new technologies to open up new frontiers for discovery and to enable costeffective reuse of computer infrastructure.

Workflow and scientific data management platforms capable of handling the volume, velocity, and variety of scientific data that will very soon exceed exabytes.

Long term and sustainable national plans, avoiding unfunded mandates, for all aspects of cyber infrastructure so that scientists can plan, share and coordinate/leverage investments.

Aggressive development and development of advanced networking.

Topical data facilities (such as the BNL RHIC Atlas Computing Facility, RACF) can be applied across multiple science domains leveraging the experience and domain expertise of the scientific/technical staff and the network/computer/storage/software infrastructure.

These are explored in more detail in other BNL position papers: Exascale Data Mining for Visual Data Analytics; Facilitating Big Data Sharing between Institutions; HEP/NP data intensive, high throughput computing programs in the BNL Physics Department; and NSLSII.

BNL Big Data Sharing

Facilitating Big Data Sharing between Institutions

Dimitrios Katramatos and Dantong Yu Computational Science Center, Brookhaven National Laboratory {dkat, dtyu}@bnl.gov

Modern-day science programs are generating numerous extremely large complex datasets, expected to soon reach the scale of exabytes. Quantitative analysis of, and knowledge discovery from, these datasets can be accomplished most effectively through extensive collaborations between domain scientists and computational experts from within an institution and/or across multiple institutions. Currently, several key challenges impede scientists from efficiently and effectively sharing such datasets; For example, existing systems for moving scientific data are not adequately general nor scalable to the required degree; the costs of commercial solutions, such as cloud drives, are prohibitive above certain quota limits, and security might be inadequate. Furthermore, the solutions do not allow institutions to host their own data using their own resources, also their transfer speeds are limited by the public Internet. Institutional site networks also may have speed limitations due to their infrastructure and administrative policies.

We envisage a Big Data sharing environment that greatly can facilitate moving data from one place to another through a straightforward user interface, while taking full advantage of the capabilities of modern high-performance network to minimize transfer times. Key elements for realizing such an environment already are being put in place or soon will be. Examples include next-generation Wide Area Networks (WANs), such as ESnet [1] and Internet2 [2] with current speeds of 100 Gb/s, and production-deployed systems to reserve bandwidths for specific traffic, such as OSCARS [3]. Also, techniques are being installed to overcome bandwidth limitations within and through the perimeter of sites via special connectivity and/or by the establishment of Science Demilitarized Zones (SDMZs). There has been investment into developing experimental systems for reserving and scheduling bandwidth within institutional networks and across multiple network domains (including WAN); systems for co-scheduling network and storage resources; and communication protocols and tools for high-performance data transfers. We feel that a strategy towards developing and
deploying a Big Data sharing environment will be highly beneficial to data users because it resolves a known need while taking advantage of existing investments. Such a framework will encompass existing technologies at the hardware and software level and focus future research towards developing more complex, enhanced capabilities that will further foster effective collaborations and scientific discoveries.

An example of such an environment would be a high-performance cloud drive for Big Data. Consider a Dropbox-style drag-and-drop service for moving and sharing files, but for files of such large sizes that would choke existing cloud drives, since the latter are meant for data volumes of a few tens to a few hundreds of megabytes. Much thought is needed on how to move a petabyte of experimental data for processing and analysis for each single dataset generated by multiple experimental cases. Such a cloud drive could be a federation of data stored at several institutions and made available through an integrating service with a user-friendly interface. Under the hood, and completely transparently to the user, this system would exploit a deployed network and storage co-scheduling infrastructure to guarantee the availability of the resources for any impending data transfer while utilizing a high-performance transfer tool with a suitable protocol to transfer the data reliably. The transfer would be scheduled to occur over special network connections created for the duration of that transfer; in this way, bandwidth bottlenecks would be avoided and the movement ultimately would conclude within a time window already known to, and agreed upon by, the user.

In pursuit of the ideas presented above, we submit a few suggestions for consideration:

• There needs to be continued and increased investment in the research and development of software that exploits the principles of Software Defined Networking (SDN) for scheduling/reserving network resources. It also should take advantage of emerging technologies, such as the OpenFlow protocol, Remote Direct Memory Access (RDMA), and RDMA over Converged Ethernet (RoCE).

• Further investments should be made in the research on and development of software that ensures easy data sharing on an unprecedented scale. An example for this would be a federated cloud drive infrastructure.

• Finally, due to the highly distributed nature of the data sharing environment, there is an evident need for the formation of a body, such as an interest group and/or a steering committee that would coordinate the work of multiple institutions. Such a committee would develop best practices, rules, and regulations further to encourage and facilitate research and development as well as the deployment and testing of the necessary software infrastructure and tools needed to implement a Big Data sharing environment.

About the authors:

Drs. Katramatos and Yu have a combined 25 years of experience in big data handling as well as in research and development of high-performance network scheduling and data transfer software, first as the staff members of the RHIC-ATLAS Computing Facility (RACF) staff and subsequently as staff researchers with the Computational Science Center (CSC) at Brookhaven National Laboratory (BNL). Dr. Yu is the group leader of the computer science group at CSC and is the PI or Co-PI of several DOE ASCR-funded networking projects such as TeraPaths, StorNet, and FTP100. Dr. Katramatos is the lead developer of TeraPaths and the PI of the Virtual Network on Demand (VNoD) project. He has participated in several other related projects such as StorNet.

References:

2. Internet2: http://www.internet2.edu

https://semanticommunity.info/Data_Science/Data_Science_for_the_National_Big_Data_R_and_D_Initiative
Updated: Sun, 14 Jul 2019 10:27:14 GMT
Powered by mindtouch™ 41
3. On Demand Secure Circuits and Advance Reservation System (OSCARS): [http://www.es.net/services/oscars/](http://www.es.net/services/oscars/)

**BNL-HEP-NP-Wenaus**

Commenting on behalf of the HEP/NP data intensive, high throughput computing programs in the BNL Physics Department.

Name, credentials, organization Torre Wenaus, Staff scientist, BNL Physics Department Distributed Computing Co-Coordinator for the ATLAS experiment at the LHC Software and Computing Manager for the US ATLAS Operations Program PanDA Distributed Workload Management System Project Co-Leader and Developer

Contact information [wenaus@bnl.gov](mailto:wenaus@bnl.gov) 631-344-4755

Experience working with big data and your role in the big data innovation ecosystem

RHIC ushered in high throughput, data intensive computing as a BNL capability 15 years ago, and LHC ATLAS computing at BNL has built on that over the past decade. The LHC is the largest scientific enterprise worldwide in data intensive computing. BNL’s ATLAS Tier 1 Center is the largest LHC computing center outside CERN. The BNL Tier 1 is part of the RHIC and ATLAS Computing Facility (RACF), a unified facility delivering leveraged, cost-effective computing to both programs, as well as many others in BNL Physics. BNL also provides the PanDA (Production and Distributed Analysis) system that orchestrates ATLAS’ data intensive computing at RACF and around the world. These capabilities make BNL one of the largest Big Data / High Throughput Computing resources and expertise pools in US science.

RACF consists of about 30 scientists and IT professionals; the PAS group consists of about 10. RACF runs over 40,000 concurrent jobs in steady state. Storage is O(100PB), about 25% of that on disk. Networking is at Terabit scale. Monthly averaged transfer rates are up to 800 MB/s between BNL and the rest of ATLAS; peaks are up to five times higher.

PanDA runs 150-200k jobs concurrently at well over 100 sites around the world, close to a million jobs a day, with about 1400 users. The system has reached Exascale in data volume processed: about 1.3 Exabytes were processed by PanDA in 2013. The ATLAS data set emerging from the first LHC run is currently about 150 PB; Run 2 (2015-2018) will accrue about six times the raw data volume of Run 1.

In its ATLAS and RHIC computing BNL is meeting challenges not only in terms of capability and scale but also in the cost effectiveness needed, in the prevailing environment of flat-flat budgets, with data processing demands growing by an order of magnitude over the next decade. BNL is a leader in developing cutting edge cost effective solutions. Examples include highly distributed federated storage serving scientists at dozens of sites across the US and globally; systems to leverage commercial clouds in a cost-effective and transparent way for data-intensive processing, interfaced to peta-scale storage; intelligent, high performance networking via ESnet, and extending intelligent network awareness to the application level; innovations in fine-grained processing workflows to utilize transient opportunistic resources and powerful networking fully, economizing on storage; and highly scalable and accessible data storage via object stores. We are extending data intensive computing to new platforms and new resources opportunistically -- high performance computing (HPCs) as well as clouds. Efficient use of storage at massive scales is an objective common to many of these initiatives – data storage is the largest cost component for data intensive computing today.

Comments and suggestions on the Visions and Priority Actions document

[https://semanticommunity.info/Data_Science/Data_Science_for_the_National_Big_Data_R_and_D_Initiative](https://semanticommunity.info/Data_Science/Data_Science_for_the_National_Big_Data_R_and_D_Initiative)  
Updated: Sun, 14 Jul 2019 10:27:14 GMT  
Powered by [mindtouch](https://mindtouch.com)
We find the high impact frontier of big data computing to be in exploiting the synergies between 1) powerful intelligent networks, 2) intelligent data-aware distributed workload management, and 3) highly scalable and accessible data stores to build agile, cost-efficient data intensive computing capabilities that can enable a scientific enterprise to leverage widely disparate distributed resources in a manageable way, making possible scales of computation well beyond what conventional approaches allow. R&D and other investments in these areas could be game-changing in giving increasingly data intensive scientific domains an accessible on-ramp to massively scaled Big Data computing.

Why you feel your contribution/ideas should be included in the strategic plan

In BNL’s HEP/NP computing program we’ve been experts at data intensive high throughput computing for 15 years, thanks to RHIC and ATLAS. ATLAS’ unprecedented scale and BNL’s leading roles in computing place BNL in the vanguard of large scale data intensive scientific computing, in the US and globally. While other science domains are relatively new to or only approaching truly Big Data computing, ATLAS’ early ascent has made BNL an expert today. BNL leverages this capability to benefit the wider HEP, NP and scientific computing communities, and is also seeking to expand to new fields such as BES and biology, and identify new ways in which its capabilities can support and advance scientific computing in the US.

BNL Integrated BigData Mining and Visual Analytics

Exascale Data Mining for Visual Data Analytics

Dr. Wei Xu, Dr. Dantong Yu, and Dr. Shinjae Yoo and, CSC BNL {xuw, dtyu, sjyoo}@bnl.gov “

Big data” from experiments, simulations, the literature, and social networks offer significant challenges of Variety, Volume, and Velocity (3-Vs) in data management and sharing, data mining, and knowledge discovery, and current computing paradigms. To resolve these programs, a holistic approach must be co-designed from the low-level hardware accelerated computing, to the top-level application-driven knowledge discovery and human-computer interaction. Large experiments facilities, for example, NSLS-II, LCLS,, along with new sequencing technology innovations system biology, healthcare generate petabytes/exabytes of image data, simulation output, and text data. Heterogeneous data must be reconstructed, analyzed, mined, integrated, to enable large scale knowledge discovery. Status-quo data analysis only considers one technique with one parameter setting, applied in a single step. Scientific research is overwhelmed by the sheer volume of data and their complexity, as measured in three aspects (3-Vs): Variety, volume and velocity. Visual analytics offers a convenient way to involve the user in problem-solving through visualization. For a large ecosystem body of data, this approach would be extraordinarily helpful in extending human experience/intelligence, enabling cognitive operations, enhancing the hypothesis generation, and testing them. Using data mining techniques would facilitate the detection of anomalies, causality or correlation in data, and eventually decision making via a human factor-aware visual analytic user interface. Adopting the four areas described as follows will greatly improve the analysis and understanding of data ecosystems:

Area 1: Interactive efficient parameter tuning assisted by visual data analytics

Nearly all the data mining and machine learning methods have parameters that must be tuned precisely; inappropriately setting the parameters will lead to misinterpreting data. With the 3-V challenges, it is very difficult to find an optimal parameter setting to fit more than one dataset. Therefore, we need 1) a learning model that simultaneously will identify
multiple parameter settings, and 2) a system wherein if a data analyst selects and tunes the specific parameter settings, then given the current models, the input from domain experts, in terms of parameter tuning and setting can be efficiently and progressively incorporated into new models.

**Area 2: Approximated online adaptive learning**

It is highly desirable to obtain the approximate outcomes rapidly for processing the data-driven learning models. Obtaining the approximate results quickly then can subsequently steer the user to more expensive, complex, yet comprehensive analysis. Especially, online streaming learning methods are among those of the greatest interest due to their one-pass learning capabilities. On top of learning from online streaming data, the analyst can provide feedbacks based on the intermediate results to the learning model which subsequently can adapt to these feedbacks (in other words, adaptive learning). To cope with the 3V challenges, especially high velocity, visual analytics can significantly enhance approximated online adaptive learning by incorporating domain experts into the learning loop.

**Area 3: Robust un-/semi-supervised learning**

It is difficult to annotate a large volume of data during supervised learning due to the excessive cost of human involvement. Unsupervised or semi-supervised learning becomes critical to visual data analytics. However, compared to supervised learning, these two models tend to be highly sensitive to tune parameter when there is a lack of appropriate annotations and users’ feedback. Even with incomplete data analysis that only have partial results available, semi-supervised learning models can effectively help analysts in finding targets progressively and quickly. We recommend robust unsupervised and supervised learning techniques to capture recurrent data events and unseen patterns easily and effectively that might be candidates for annotation and parameter tuning. Visual data analysis is the bridge to integrate these two types of learning.

**Area 4: Algorithm Acceleration with state-of-the-art hardware**

In response to the anticipated high volume of data generated by advanced observational platforms and fine-grain model simulations, the models and methodology are often data-/computing intensive. It is highly desirable to extensively use state-of-the-art computational technologies, such as high-performance parallel processing, GPGPU and the Intel Phi coprocessor to accelerate algorithm modules, building blocking, dependency libraries and also to cope with the challenges of data volume and velocity.

In pursuit of the concepts discussed above, we offer a few suggestions for consideration:

- There needs to be continued- and increased- investment in the research on and development of basic algorithm research in data mining and knowledge discovery.
- Data mining and visualization needs to be integrated to cross-validate the results of data mining by human involvement. However, directly using existing visual analytic techniques can hardly satisfy the demands for analyses. Customization is needed when developing visual analytics tool for specific applications. The development usually includes the collaboration of experts from related multiple disciplines. There are many cutting-edge publications as showcases of the customization process with current visual analytics techniques. Undoubtedly, more investment is needed to transfer these research outcomes into practical system and foster new analysis algorithm to be developed.
- On-line data sampling and streaming analysis will be critical as data must be analyzed before most of it is archived and analysis of it becomes difficult.
Data integration and the analysis of fused data must be supported to reveal a comprehensive picture of science phenomena as the same data samples are measured, and evaluated under various circumstances.

Wei Xu received her Ph.D. degree in Computer Science from Stony Brook University, USA in 2012. She joined Brookhaven National Lab (BNL) in 2013. Her research interests include medical imaging, tomography, visualization, visual analytics, GPGPUs, machine learning and workflow systems. She has published papers in leading technical journals and conferences and served as committee member and reviewers for top medical imaging and visualization journals and conferences.

Dantong Yu is a research scientist and group leader at the computational science center. His research interests include HPC, data mining and science-driven knowledge discovery. He has published more than 50 papers in leading CS journal and conference proceedings.

Shinjae Yoo is a research scientist and group leader at the computational science center. His research interests include bioinformatics, data mining and machine learning. He has published work on data mining and machine learning.

References

- H. Huang, S. Yoo, H. Qin, and D. Yu, Physics-based Anomaly Detection Defined on Manifold Space, Accepted to ACM Transactions on Knowledge Discovery from Data, 2014 (TKDD).

BNL-NSLS2

National Synchrotron Light Source II response to NITRD RFI for National Big Data R&D Initiative

Paul Zschack, pszschack@bnl.gov, (631) 344-8703 Brookhaven National Laboratory, P.O. Box 5000, Upton, NY 119735000

National Synchrotron Light Source II (NSLS-II) is a state-of-the-art 3 GeV electron storage ring designed to deliver world-leading photon intensity and brightness over a wide spectral range including infrared, ultraviolet, and x-ray. The facility is operated as a DOE Office of Science User Facility for the benefit of widely diverse scientific user communities. When fully developed, the facility will host up to 4000 users each year, and will include over 60 simultaneously operating beamlines, each providing different advanced capabilities. NSLS-II will offer researchers from academia, industry, and national laboratories new ways to study material properties and functions with nanoscale resolution and exquisite sensitivity by providing state-of-the-art capabilities for x-ray imaging, scattering, and spectroscopy. Users will come from all 50 states (and from around the world) to study problems across many scientific disciplines. This research is sponsored through various channels, including NIH, NSF, DOE and others.

High brightness and high flux available at many of the NSLS-II beamlines coupled with advanced high-speed megapixel detectors will drive much higher data rates and volumes. Indeed, all synchrotron disciplines are increasingly using multi-
element detectors to efficiently measure and record the results of the experiment, and certain techniques often utilize detectors that can produce multi-megapixel images with frame-rates in the kilohertz range. The volume of data associated with a single experiment may exceed several tens of terabytes, and in the next few years, the facility may produce nearly 20 petabytes annually.

The broad uses of new megapixel detectors together with new, advanced techniques drive a new paradigm for data processing and analysis at NSLS-II and at synchrotron sources worldwide. Individual users are often unable to write the sophisticated algorithms required to process, analyze, visualize, interpret or otherwise extract the important information from raw detector data. Multi-dimensional data sets are intrinsically more difficult and often require specialized data reduction and manipulation algorithms and routines not readily available to individual users.

Building and extending an infrastructure (hardware and software) to improve access for users who are generally not expert in Big Data problems, computing, visualization, etc… is essential to realize the full potential of NSLS-II. While users of the facility are often leading in their respective disciplines, many will not have the necessary resources at their home institution to productively develop knowledge from the data they collect. Further, the ability for users to remotely access, manipulate, analyze and visualize their data will be essential.

Many complex problems span orders of magnitude in spatial or temporal extent. These multi-scale problems are well suited to Big Data approaches that are well adapted to work with heterogeneous data sources. These data sets, together with data collected at one or more synchrotron beamline may be combined to form a more complete picture of a sample or process. Furthermore, users are increasingly applying widely varying and complimentary analytical techniques to study complex materials or processes. Combining these other measurements with data collected at the light source provides an opportunity for a comprehensive understanding and suggests a new pathway for discovery. Tools to facilitate these complex and transformative approaches are needed, as well as research into quantifying uncertainties when combining different data sets.

As synchrotron experiments become more and more sophisticated, there exists an increasingly greater need for adequate theoretical modelling and simulations. Many experimental results have a significant dependence on coherent properties of the beam as determined by beamline optics and beam defining apertures. Therefore, better understanding of how a complete experiment works including the effects of the source and optics is important for correct interpretation of the experimental data. Parameters of the beamline (metadata) are collected together with detector data and can be used for simulation and modelling to greatly enhance the NSLS-II capabilities in experimental planning and design. Improved interpretations of experimental data will provide research results with a greater impact on the research activities in variety of scientific disciplines.

Finally, the value of archiving users’ data must be weighed against the costs associated with ensuring the long term sustainability, access, and curation of users’ data. Without a detailed process knowledge of the sample, the raw data from any users’ experiment does not have particularly significant value to other researchers. The experimental results are largely determined by the particular sample and its preparation or processing. On the other hand, as new data analysis methods evolve, older data may still hold considerable value for new discovery. Well-articulated policies on storage of large data sets should be developed and appropriately resourced.
Response to Request for Input (RFI)-National Big Data R&D Initiative

Center for Nanoscale Materials Argonne National Laboratory 9700 S. Cass Ave., Argonne, IL 60439 USA

Contacts:

Ian McNulty X-ray Microscopy Group Leader, Senior Scientist mcnulty@anl.gov, 640-252-2882
Katie Carrado Gregar CNM User & Outreach Programs Manager kcarrado@anl.gov, 630-252-7968

Big-Data experience: Significant quantities of scientific data are being generated at the Center for Nanoscale Materials by both experimental and theoretical approaches. The highest data rates reach 1 TB on time scales of an hour; several activities at CNM generate this scale of data within a few days. We anticipate that these data rates will increase by an order of magnitude over the next five years. Nearly all of this data requires extensive analysis, and in an increasing number of cases, high performance computing resources, to extract scientifically useful information from it. In addition to producing Big Data on a sustained basis, CNM is actively planning how to manage its current data explosion and correspondingly increasing needs over the next few years. We coorganized a workshop addressing several of these issues at the 2013 CNM/EMC/APS User Meeting titled, "Driving Discovery: Visualization, Data Management, and Workflow Techniques", and participated in a meeting held on 25-26 October, 2014 in Melbourne, Australia, titled "Big Data X-ray Microscopy Workshop". We are participants in a joint ASCR/BES-funded project as well as ongoing Big Data-related collaborations with staff at Lawrence Berkeley National Laboratory and Brookhaven National Laboratory.

We addressed the NITRD submission questions as follows:

1. What are the gaps that are not addressed in the Visions and Priority Actions document?

Response: The principal gap in the Visions and Priority Actions document that impacts our community now, and that is expected to impact it into the future, is analysis between acquisition of data and its reduction into publishable and usable form. While far from ideal, we are better at storing the raw data as it is generated, and organizing processed and reduced data into a form suitable for dissemination, then we are at keeping up with the analysis necessary to process and reduce the raw data (the "Analysis Bottleneck").

2. From an interagency perspective, what do you think are the most high impact ideas at the frontiers of big data research and development?

Response: The National User Facilities such as the Nanoscale Science Research Centers and Light Sources have a significant Big Data problem: a growing fraction of the prodigious amount of data they produce is not getting analyzed. The primary sources of this flood of data are multispectral, megapixel detectors with high frame-rates (up to 10 Kframes/s), substantially brighter (100-200x) x-ray sources, and increasingly larger-scale (up to petascale) computational capabilities being developed to analyze and simulate more complex (multidimensional) problems and atomistic (more than 1 million atom) systems. From a facility utilization standpoint it is highly inefficient to take days to months to analyze a data set acquired in minutes. Guiding the course of a research program or making critical decisions on the time scale of the data acquisition becomes impossible under these conditions. Solutions to the Analysis Bottleneck- new analysis
approaches, paradigms and infrastructure for keeping up with the data being collected - will have the highest impact on our community, and provide the best possible utilization of the considerable investment in our National User Facilities such as the CNM.

3. What new research, education, and/or infrastructure investments do you think will be game-changing for the big data innovation ecosystem?

Response: The game-changer for our community would be a concerted investment into infrastructure and methods to resolve the Analysis Bottleneck. This, more than any other investment, will create the best possible opportunity space for scientific discovery. This requires investment well beyond than a brute-force approach, e.g. just adding more computational capability, data storage space, and faster computer networks. We believe a more effective approach for our community, in addition to hardware investments, would be to engender greater awareness among our users and staff through workshops and seminars about the opportunities and challenges presented by this bottleneck, to support R&D that addresses these opportunities and challenges through more effective datasharing, algorithmic, and analysis methods, and to encourage partnerships with other laboratories and groups facing similar challenges so as to address them together.

4. How can the federal government most effectively enable new partnerships, particularly those that cross sectors or domains?

Response: The most effective ways that government can enable new partnerships in our user facility environment is to promote awareness about the challenges and opportunities of Big Data, to support longer term as well as pilot grants via proposal calls for development of education, new approaches and infrastructure to manage Big Data more effectively, and to encourage inter-laboratory (e.g., between CNM and the Molecular Foundry) as well as cross-divisional collaboration (e.g. between DOE-BES and DOEASCR).

5. A short explanation of why you feel your contribution/ideas should be included in the strategic plan.

Response: As a National User Facility we represent the needs of a broad community of scientists that will benefit from improvements in managing Big Data, leading to greater scientific productivity and facility utilization.

Additional information: We surveyed the CNM staff and User Community to estimate current and future data generation rates, and to consider what aspects of our local Big Data environment pose the greatest bottlenecks to scientific productivity at CNM. Table 1 summarize our findings. A 1-5 scale was used to rate the significance of the data bottlenecks, where 1 is most and 5 is least significant.

Table 1. Results of data generation rates and bottlenecks at CNM.
Columbia

Response From The Data Science Institute, Columbia University

Director, Kathleen McKeown, Rothschild Prof. of Computer Science, kathy@cs.columbia.edu

Executive Committee Members Garud Iyengar, Prof. and Chair of Industrial Engineering and Operations Research Paul Sajda, Prof. of Biomedical Engineering, Radiology (Physics) and Electrical Engineering

The Data Science Institute at Columbia University is a broad interdisciplinary Institute with five Centers in application areas of data science and one Center focusing on the foundations of data science. Our mission is the development of technology that can exploit the massive amounts of data available today to help solve society’s most challenging problems. We bring together researchers on interdisciplinary teams to enable the use of data science in disciplines from journalism to science. We have launched a certification in data science for working professionals, an MS in data science and are working on a PhD program in data science.

Amidst the arrival of massive, streaming, and heterogeneous datasets, we argue that funders must retain a focus on the science of big data. In order to maintain breakthroughs in the impact derived from big datasets, it is critical that fundamental progress continue on representation, algorithms, and statistical models surrounding the use of data. Without these, big data loses its potential. Cross agency efforts can accelerate this progress as the same methods may be developed for radically different big data applications. For example, a major innovation in novel probabilistic models, approximate inference algorithms, or theoretical guarantees can make a huge impact across many domains. By requiring innovation in data science techniques as well as in the application, cross agency efforts can accelerate research and the potential in the application itself. This approach pairs data scientists with experts in disciplines on all projects.

In the specific research challenges we present below, researchers from different disciplines must be able to work together to develop solutions. Strategies for bridging disciplinary boundaries and building knowledge drawn from diverse teams must be developed for both educational and research practice.

Specific Research Challenges
**Health:** In health analytics the lack of shared, linked, meaningful and interoperable datasets is a key obstacle to progress. Privacy of personal health information is critical and approaches that enable privacy preserving summaries over both clinical and genetic data are needed. Given that data is streaming over time, is heterogeneous, and includes unstructured text, images and numerics, new techniques are needed that can align and anchor multiple data types automatically to identify time-resolved salient events of medical/clinical significance. New research is also required to leverage the potential of wearable sensors as a medical/health data source and its integration with hospital and physician medical care records. This area is likely to result in a “data-driven revolution” into the way we diagnosis and treat many of our societies’ most chronic and costly diseases, including obesity, diabetes and heart disease. The technical challenges are so broad, and the impact to the average American is so significant, that a multi-agency effort is warranted (NIH, NSF, NIST, DoD).

**Disaster, risk and resilience:** Several agencies are separately focused on different activities to improve disaster response, resilience, and associated risk management (DHS, DARPA, NIH, NIST). Similarly, disparate research communities are working on various aspects of risk management of disasters: analysis of social media for crisis informatics; simulation and computational modeling; developing resilient infrastructure; and understanding of social behavior in the context of epidemics and other public health crises. Coordination among agencies and research groups would lead to a significant impact. Further advantages are to be had by encouraging the development of methods that are able to jointly exploit different existing data sets: FEMA datasets, HAZUS models, hand-built taxonomies for specific communities, NOAA weather data, economic models; and new datasets, such as databases of global supply chains with precise geographic locations of manufacturing and distribution facilities so that risks associated with natural disasters could be addressed. There is a need to make all available data publically accessible in a format that would enable cross-data set analysis.

**Natural and applied science:** Research on the study of the ocean and our forests offers tremendous potential for the understanding of climate change. Using heterogeneous data collected by NOAA, NASA and NODC may help to predict the impact of the melting Arctic ice on the North Atlantic. Research that connects data science with material genomics, astronomy and physics is also key. Materials performance, for example, is the bottleneck in critical technologies such as sustainable energy (e.g., solar energy production). Big Data approaches are well suited to this but there have been virtually no approaches along these lines to date. Bringing scientists together with data scientists who have expertise in optimization has promise to yield computationally efficient solutions. DOE, NASA, and NSF can support such efforts.

**Social science:** Data-driven research is also needed to support policy decisions that require interaction between government and academia. New York City DOH’s research on how the lunch program affects child obesity is a good example of the use of city data for public health. Estimating populations at risk for infectious diseases such as HIV/AIDS, Hepatitis C and estimating the effects of risk behavior combined with social structure and effective campaign and communication strategies to increase awareness of programs such as flu vaccines. Research challenges in all of these areas include: study design, data collection, processing and analysis, and policy analysis through simulation studies based on findings from field data analysis to support the decision making process and they cross agencies such as NIH and NSF.

**Urban Data Science and Decision Support:** Given the explosion in urban sensors and data sources, advances in data science tools and urban simulation platforms, and domain expertise across the engineering, natural and social sciences, we can now accelerate knowledge in the field of multi-scale urban dynamics and provide decision support tools for charting sustainable city futures. Funding needs to focus on scientific and technical advances within, and at the
intersection of, urban sensing (from a wide variety of novel sources including advanced infrastructure monitoring networks, social networks, mobile data and transactions), computational urban simulation and modeling, and optimization and decision support tools.

Mechanisms for Funding Education

New mechanisms for funding are needed to support the educational enterprise around big data. A funding focus on big data as it relates to an entire ecosystem, as opposed to a specific research project, would enable major advances. For training, PhD fellowships that place a cohort of students who excel in the foundations of data science together with cohorts of students interested in exploiting the big datasets in their discipline. This would enable the education of many more students than is possible with investigator focused research projects. The creation of national laboratories, each focusing on one or more disciplines in combination with data science, would also accelerate research. Finally, given the need to immerse oneself in both technology and data from different disciplines, longer periods of time before becoming a faculty member or joining industry is needed. Time to learn the vocabulary of other disciplines and the importance of different data is essential. Thus, we also advocate for the establishment of data science postdoctoral fellowships.

CSC

RFI Response for National Science Foundation (NSF) – National Big Data RFI Initiative

1. Who you are—name, credentials, organization, contact information; your experience working with big data and your role in the big data innovation ecosystem

CSC is a global leader of next-generation IT services and solutions. Our mission is to enable superior returns on clients’ technology investments through best-in-class industry solutions, domain expertise and global scale. CSC brings our broad and deep expertise, in concert with a proven track record of extracting, processing and implementing IT and data solutions across a variety of clients. CSC has successfully delivered a multitude of complex data integration projects for federal and state government agencies, hospitals, and global healthcare institutions (e.g. United Kingdom National Health Service program) and Defense and Intel customers.

Point of Contact (POC): Yvonne Chaplin Big Data & Analytics Brand Manager, Computer Sciences Corporation 3170 Fairview Park Drive Falls Church, VA 22042 C: 571-294-9875 email: ychaplin@csc.com

2. What are the gaps that are not addressed in the Visions and Priority Actions document?

1) Organizational Transformation: To fully realize the value of Big Data, organizations will need to take action based on the insights that the data provide. For example, data analysis with the purpose of enhancing the customer experience may lead to recommended changes in business process, data sources, infrastructure, architecture, and organizational structures.

2) Use Cases: We recommend the National Big Data Initiative identify specific use cases to which big data solutions may provide remedies. The focus on specific problems to solve would ensure practical use – rather than building a solution in search of a problem.
3) Data Governance: Data governance within and across organizations is key to ensuring long-term sustainability and access. Data governance agreements that articulate who can access the data at what level, privacy considerations, and de-identifying personal information will help to break down the “stovepipes” between organizations in a secure manner.

3. From an interagency perspective, what do you think are the most high impact ideas at the frontiers of Big Data research and development?

1) Public safety: Enhanced forecasting of the extent and impact of natural and man-made disasters and improved status will continue to help optimize the deployment of resources to support and protect the population.

2) Electronic health records (EHRs): EHRs coupled with new analytics tools present an opportunity to mine information for the most effective outcomes across large populations.

Using carefully de-identified information, researchers can look for statistically valid trends and provide assessments based upon true quality of care.

3) Health sensors: Used in the hospital or home, these sensors can provide continuous monitoring of key biochemical markers, performing real time analysis on the data as it streams from individual high-risk patients to a HIPAA-compliant analysis system. By providing alerts to a health anomaly or pending critical event, sensors have the potential to extend and improve the quality of millions of citizens’ lives.

4) Education: In-depth tracking and analysis of on-line student learning activities – with fine grained analysis down to the level of mouse clicks – can help researchers to ascertain how students learn and the approaches that can be used to improve learning. This analysis can be done across thousands of students and can inform the development of courses and teaching approaches to reflect the information gleaned from the large scale analysis.

4. What new research, education, and/or infrastructure investments do you think will be game-changing for the big data innovation ecosystem?

1) Research: We recommend continued investment and support of Open Source projects to promote speed and collaboration among a community of Big Data innovators. In addition, we are seeing an explosion of social media data that is publicly available. Enterprises and Government have incredible opportunities to fuse their own data with this public social media data to draw intelligence.

2) Education and training: Data scientists will require a multi-disciplinary skill-set combining business, mathematics, statistics and engineering. Data scientists are not typically technologists, and often have great difficulty manipulating the complex technologies that span the big data spectrum. R&D should focus on bridging this gap by developing tools (one example is open source Amino) to simplify this complexity for the analyst.

3) Infrastructure: We recommend that infrastructure R&D address delivery of information through varied, multiple, and concurrent channels and mobile devices.

5. How can the federal government most effectively enable new partnerships, particularly those that cross sectors or domains?
Data governance agreements and security standards will be critical in enabling disparate parties to form partnerships, providing a degree of assurance that privacy and sensitive data will be appropriately accessed and protected.

6. A short explanation of why you feel your contribution/ideas should be included in the strategic plan (examples where appropriate)

CSC’s response and recommendations are based on practical “real-world” experience of applying Big Data strategies and technologies and developing solutions for multiple clients in the public and private sectors. Examples are briefly cited in our response to the first question above.

Dinov_UMich

I. RFI Responder: Ivo D. Dinov, PhD, Associate Professor and Director, Statistics Online Computational Resource (SOCR), University of Michigan, Ann Arbor, MI 48109 (734) 615-5087, http://umich.edu/~dinov, statistics@umich.edu.

II. Responder Big Data Experience: In my ongoing work, I have done research 1,2, designed innovative data analytics services 3,4, developed distributed software tools 5,6, and integrated the computational, theoretical and applied components of Big Data Science and Infrastructure into the graduate curriculum 7,8. As an example, a recent article, “The perfect neuroimaging-genetics-computation storm: collision of petabytes of data, millions of hardware devices and thousands of software tools” 9, illustrated that the Kryder’s law for exponential increase of the volume of data is real. Although Moore’s law for exponential increase of computational power (transistor capacity) and Kryder’s laws indicate similar exponential expansion over time, the rate of increase of data volume is significantly higher than our ability to manage, process and interpret that data we collect. Relative to the Moore’s law, the exponential parameter in the Kryder’s law for data increase is bigger.

III. Comments on Initial Framework: Although the National Big Data R&D Initiative Framework 10 pinpoints some significant Big Data challenges and opportunities, it falls short of what is necessary to ensure effective, coordinated, agile and sustainable model for proactively addressing the needs for managing the avalanche of data in all aspects of daily living. The 3 most notable shortcomings of this initial plan include crowd-source engagement, agile development and training sustainability. Current and future innovations in Big Data analytics are most likely going to come from disparate resources, small-group initiatives, open-source/open-science community and truly transdisciplinary interactions, not from Big-Business, Big-Academy, or Big-Government. We need a concerted effort to harness the ingenuity of the broader community using large-number of smaller-budget, rapid-development, high-risk, product-oriented projects (including funding mechanisms). The era of Big Data analytics is here and we need to realize continuous-development, rapid agile prototyping, and experience-based (e.g., evidence-based) redesign are the new normal for all innovation, including basic science, computational modeling, applied research or studies of system complexity.

IV. Gaps and Barriers: (1) Gaps in technological skills and expertise to conduct high-throughput Big Data analysis, (2) lack of open, interactive and interoperable learning resources, (3) challenges of sharing tools, materials and activities across different disciplines, (4) considerable discipline-specific knowledge boundaries, (5) challenges of fusion of qualitative (e.g., study design) and quantitative (e.g., analytic modeling) Big Data methods, (6) availability of integrated and interoperable Big Data resources. For example, a SOCR study is currently looking at healthcare data (e.g., CMS, CMS, https://semanticommunity.info/Data_Science/Data_Science_for_the_National_Big_Data_R_and_D_Initiative
Updated: Sun, 14 Jul 2019 10:27:14 GMT
Powered by mindtouch
WHO), demographic trends (e.g., Census), economics factors (e.g., BLS) and social trends (e.g., Web-traffic, Tweeter) aiming to determine the associations and organization of data elements from many hundreds of data elements in the presence of incongruent sampling, complex, incomplete and heterogeneous data from different sources.

V. High-Impact Ideas: Decisively, there are 4 specific and complementary funding directions that could significantly impact the process of extracting information from Big Data, translating that information to knowledge, which can in turn guide our actions: (1) Enforce (for real) open-science principles; (2) Engage and actively participate (e.g., fund) nontraditional high-risk/high-potential-impact studies; (3) Adapt to rapid agile development, testing, redesign, productization and utilization of data, tools, services and architecture; (4) Redesign the data science curricula (from high-school to doctoral level training). Big Data is incredibly powerful, but its Achilles hill is time. Its value is in the moment and its importance decreases exponentially with time 11, which makes the effective and rapid response to collected data critically important.

VI. Investment Targets: We need to demand open-science, reduce barriers to sharing of data, protocols, and tools, cast a wider funding net and diversify the scope of research and development efforts, and finally, significantly overhaul Data Science education.

VII. Partnerships: Enabling new public-private-government partnerships is easier said than done. There are many complexities, conflicts of interests, and divergent visions that affect the agreements establishing effective cooperations between organizations and independent institutions. Long- and short-term goals can collide and misalignment of interests, bottomlines and scope of impact can bifurcate causing frictions and difficulties in managing extremely diverse partnerships. Albeit not forced, partnerships should be encouraged and facilitated as necessary. The best scientific discoveries are advanced and popularized not by regulations, but by loose interactions, open-sharing and collaboration.

VIII. Justification of Contribution: There are 2 ways to deal with the influx of significant disruptive technologies and – reactive response (passive) or proactive action. Government institutions and regulators, funding agencies and organizations involved in generating, aggregating, processing, analyzing, interpreting or managing large, incongruent, heterogeneous and complex data may choose to lead or follow the Big Data revolution. As engaged citizens, energetic researchers and diligent parents, we need to ensure our National priorities position us well for riding the Big Data wave, as opposed to training in its wake. The National Big Data R&D Initiative and the Networking and Information Technology Research and Development group are in a powerful position to gently guide, actively engage, and significantly stimulate future innovation in Big Data science, engineering, applications, knowledge and action.

1 http://dx.doi.org/10.3389/fninf.2014.00041
2 http://www.socr.umich.edu/people/din...lications.html
3 http://socr.umich.edu/HTML5/
4 http://socr.umich.edu/HTML5/BrainViewer/
5 http://pipeline.loni.usc.edu/get-started/acknowledgmentscredits/
6 http://distributome.org
Dear BDSSG Stakeholders:

My name is Edward Rescsanski and I am the founder of Executive Decision Consulting LLC (EDC) in Fairfield, CT. EDC specializes in Big Data, Analytics and Data Governance. We are a small, focused group that can help immediately. If we engage in a partnership, I will be your point of contact and the driver of this initiative. Unto that end, I will briefly profile my experience.

I have worked in the information technology field for over 20 years and I started at the bottom, eventually working my way up to owner of EDC. My background is full of storage, operating system, disaster recovery and security experience. Before creating EDC, I was even VP of Information Security at one of the largest banks in the country. Through it all, I gained experience in analytic process, product and implementation. Today, I can provide in-depth knowledge of Big Data vision, integration, and a variety of ways and vendors that connect dissimilar data in order to extract value.

- When I discuss Big Data or Analytics, I commonly refer to all of it as the Value Extraction of Data (VED). The reason is that too many of us are fighting over definitions and semantics which in turn takes away from our ability to innovate. Big Data, Analytics, BI, Data Warehouses, and Data Mining all have a single goal and that is to extract value from data. Most Big Data experts will say that there are 5 V’s (Volume, Velocity, Variety, Veracity, and Value). To me, Volume, Velocity, Variety and Veracity are descriptors of data. Value is the payoff that we, as humans and organizations, receive as a result of combining the four descriptor types. The exercise of defining a common vision comes before all else.

- The Governance of a Big Data ecosystem comes next and is critical as it requires the policy documentation (laws to be followed), the process documentation (methods of staying in compliance with policy) and a universal organizational training and buy-in. EDC provides that documentation and regularly bridges organizational silos to create innovation.

- After the Governance Framework is established, ecosystem templates must be built. These templates will include cost of goods for hardware, software, maintenance and services. EDC utilizes a framework called the IT Triad. It is a process that perfectly aligns technology with an organizational purpose. Remember, Big Data is still data and unto that end, it must be stored, protected and transported just like any other application, though in this case the scale can be enormous. The building, scaling, and protecting of your Big Data ecosystem requires this step to be flawless.

- Defining data types and locations of that data will provide answers as to which connectors must be used, security levels of people accessing the data, and the “time classification” of the query. Time classification relates to real-time, near-time or not time sensitive. An example of each would be:
Real Time – Government receives warning that a terrorist group will strike U.S. soil within days. Big Data helps determine legitimacy, location, time, type and persons involved.

Near Time – Lyme disease spread and cure rates. Every year it grows and more people suffer. Is chronic Lyme real? Long term treatment? CDC doesn’t provide much of any acknowledgment on true numbers. As next summer gets closer, the “Near Time” solution fits well.

Not Time Sensitive – Astronomers discover a new planet, possibly habitable. Let’s dig in and invite as many qualified people to see the data, add to data and contribute.

In conclusion, this can’t possibly fit on two pages…Your Big Data partners must have an open mind and if you choose an independent, like EDC, you get an agnostic approach, leveraging experience and process from many vendors and diverse situations. EDC provides accountability at a level higher than any other organization.

How can we discover the potential of Big Data? By focusing on integrity and accountability. If you want a game changer, place diverse but like-minded people together and lead them to achieve a common goal. That goal is transcendence.

Sincerely, Edward Rescsanski

ESIP

Introduction

This response to the NITRD RFI and draft Big Data R&D vision is submitted on behalf of the authors of a workshop report on data science priorities in the Earth and environmental sciences. In the report, the participants identified the need for comprehensive, cross-agency efforts to address existing and looming challenges related to the creation, management, use and re-use of agency sponsored scientific data. The participants strongly urge the following initiatives be put in motion related to science data challenges: 1) enable the National Research Council to lead a high-level task force study of these issues to provide strategic guidance that agencies can use in planning, 2) consider the creation of a cross-agency effort to create a sustained Science Data Infrastructure (SDI). The full document, Workshop Report: Planning for a Community Study of Scientific Data Study, including an executive summary may be found at http://dx.doi.org/10.7269/P3R49NQZ. The workshop was funded by the Gordon and Betty Moore Foundation, by the National Consortium for Data Science (http://data2discovery.org), and by ESIP (Federation for Earth Science Information Partners, http://www.esipfed.org).

Genesis of the Workshop / Workshop Participants

The data science workshop arose from within the Earth science community as represented by ESIP. ESIP and its members epitomize the broad swath of Earth science communities and agencies charged with observing and monitoring the Earth, including NASA, NOAA, USGS, and EPA. ESIP formed a working group to consider the idea of a data science study and ultimately led the way in defining, organizing and leading the workshop. The participants included recognized leaders in the fields of Earth and environmental science and data science. While the workshop focused on Earth science and related domains, participants recognized that these challenges and findings were likely pervasive across the spectrum of scientific and applied research.

Summary of Needs Identified
The report identifies five categories of challenges around scientific data: 1) economic, 2) cultural, 3) data science research, 4) educational, and 5) legal, ethical, and policy.

The primary challenge relates to the economics of data. We do not know the value our data provide. Though, research is growing into how to measure both actualized and generative value of research data. Early results show significant positive returns on investment from two to twenty times the cost. Currently, data management costs are typically scraped from research funds; with the associated tension and resistance that situation creates. To create an effective, sustainable funding model the economics need to be better understood.

Cultural changes are needed to recognize and reward the value of dataset creation, management and stewardship. Data science research includes questions about data discoverability, access, interoperability, stewardship, and a possible mathematics or science of data, among others. Data stewardship must be integrated into science education, and appropriate data must be accessible for science education at all levels. Finally, legal, ethical, and policy challenges include licensing issues, achieving scientific goals of verifiability and repeatability, as well as the possible creation of an agency, council, or office to support scientific data infrastructure.

**Recommendation #1: Convene a high level, fast track study of cross-agency needs and future strategies. The study should be conducted by the National Research Council as a highly respected neutral party.**

The participants in the workshop strongly urge the creation of a high-level National Research Council–led study “to conduct an authoritative, unbiased assessment of strategic scientific data investments.” The purpose of the assessment, as described in the report, is to provide strategic guidance on improving the management of scientific data in a cost effective manner to support the national interest. In addition, the study is intended to establish or re-establish the United States as a leader in science data infrastructure. Participants also encourage the use of an approach to the study that emulates a high profile panel approach. A long, drawn out traditional study is not warranted in this case. Any proposed study would also need to review and synthesize relevant prior work in science data management and infrastructure.

**Recommendation # 2: Envisioning a National Science Data Infrastructure (SDI)**

Workshop participants had a bold vision for a sustained Science Data Infrastructure (SDI) and associated technical and cultural shifts to better enable science in the face of major challenges now and into the future. In order to maximize the return on government investments and to minimize the duplication of systems and efforts, a cross-agency entity—something more than a working group or task force—could be charged with attacking this problem. The end uses are fundamentally cross-agency, cross-disciplinary, and across temporal and spatial scales. The workshop participants recognized that science questions are increasingly complex and require inter- and trans-disciplinary approaches for conducting research. The envisioned SDI would support science by providing a common framework under which all science data would be governed. The challenge is to transcend traditional agency missions in favor of a bigger, grander vision.

**Conclusion**

The workshop participants are well placed to assess the types of challenges facing US Federal agencies in regard to data. These individuals work with the agencies and the data they generate, directly or through sponsored activities, in an ongoing basis. These individuals also work with various groups such as BRDI, CODATA, ESIP, RDA and others to
exercise leadership in these areas. We look forward to engaging with NITRD in this effort as they consider these recommendations and the others submitted in response to this RFI.

Point of Contact:

W. Christopher Lenhardt, Domain Scientist, Environmental Data Science and Systems Renaissance Computing Institute, Chapel Hill, NC, clenhardt@renci.org, 919--445--0480.

ESnet

From: Inder Monga, CTO, and Eli Dart, Senior Network Engineer, Energy Sciences Network (ESnet), Lawrence Berkeley National Lab, Berkeley, CA imonga@es.net and dart@es.net

Introduction

The Energy Sciences Network (ESnet)1, a division of Lawrence Berkeley National Laboratory2, is the Department of Energy's high-performance networking facility, engineered and optimized for big-data science. ESnet interconnects the entire national laboratory system, including its supercomputer centers and user facilities. This enables tens of thousands of scientists, at 40 DOE sites, to transfer Big Data, access remote resources, and collaborate productively via access to over 100 research and commercial networks worldwide. The richness of ESnet's global connectivity is motivated by the fact that approximately 80% of its traffic originates or terminates outside the national laboratory complex. This pattern in turn reflects the collaborative, increasingly international, nature of scientific research

Experience

ESnet was founded in 1986, soon after the creation of the global Internet. ESnet's mission has been to enable and accelerate scientific discovery by delivering unparalleled network infrastructure, capabilities, and tools. Since 1990, ESnet's traffic has increased by a factor of 10 every 48 months as seen in Figure 1, roughly double the growth rate of the commercial Internet 3.

Figure 1: ESnet accepted traffic since 1990 (PB/month)
The structure of modern science now presumes the availability of reliable, high-bandwidth, feature-rich networks for interconnecting instruments and collaborators globally. As science is now largely data-intensive, ESnet focuses its efforts on the effective support of data-intensive science. ESnet has developed its architecture, capabilities and services primarily to serve scientific users and use cases, and maintains its understanding of the needs of its constituents through a formal requirements-gathering process.

In addition, ESnet has developed a wealth of real-world experience in enabling data-intensive science. This experience is captured in the fasterdata knowledge base (fasterdata.es.net) which also includes the Science DMZ architecture, created by ESnet. Science DMZs have been widely deployed in the US, with the NSF creating programs (CC-NIE and CC-IIE) to fund the adoption of this architecture. These programs have laid the foundation for a revolution in data-intensive science in the US, and many countries around the world are following suit. ESnet also develops open-source tools for science networks, including perfSONAR, which has seen over 1200 deployments globally, and OSCARS, which received an R&D 100 award in 2013.
Data movement, networking and its importance in big-data initiatives

As we look at the science constituency worldwide, unconstrained data mobility is extremely important and many times largely ignored by ‘Big Data’ initiatives. Large data sets are often moved to facilities where they can be analyzed; moved again for secondary analysis; and eventually shared with collaborators. This pattern applies broadly, and is a hallmark of data-intensive science.

Computing has become essential in virtually every area of science -- from extreme-scale simulations to high-throughput analysis of instrument data, the use of computing in science is ubiquitous. Networking is also critical to data-intensive science – it provides for remote access, data movement, workflow integration, data integration, and collaboration. The effective use of networking technologies works as a lever, significantly increasing the value and productivity of interconnected scientific facilities, supercomputer centers, and collaborations.

Why this contribution should be accepted

We welcome an environment where scientific progress will be completely unconstrained by the physical location of instruments, people, computational resources, or data. To achieve this, we believe, future Big Data initiatives must:

- Produce friction-free network architectures, free of physical or logical flaws, which do not impact the ability of protocols to function in a correct and efficient manner.
- Support full-featured applications that have access to a deeper understanding of network construction, behavior, and expectations.
- Foster the deployment and operation of intelligent networks that can function with autonomy when needed, or through direct control by higher-level entities to deliver higher classes of service.
- Support high-performance data migration, data placement, and data access in ways that are productive for users and allow the development of a new class of data-driven scientific methods

1 ESnet web site, http://www.es.net
2 Lawrence Berkeley National Laboratory, http://www.lbl.gov
4 ESnet Requirements Review Program, http://www.es.net/requirements/

Robert Murphy is the Big Data Program Manager for General Atomics Energy and Advanced Concepts. Previously, Robert was responsible for High Performance Data and Analytics in IBM’s Software Defined Environments organization. Before IBM, Bob held positions with increasing levels of responsibility at Hewlett Packard, Silicon Graphics, Oracle, and Sun Microsystems. He has a Biomedical Engineering Degree from Purdue University.

Robert Murphy Big Data Program Manager, Energy and Advanced Concepts General Atomics 3550 General Atomics Court, San Diego, CA 92121 robert.murphy@ga.com

https://semanticommunity.info/Data_Science/Data_Science_for_the_National_Big_Data_R_and_D_Initiative
Updated: Sun, 14 Jul 2019 10:27:14 GMT
Powered by mindtouch
Comments and suggestions from General Atomics

1) Data from large-scale experiments and extreme-scale computing is expensive to produce and may be used for high-consequence applications. However, it is not the mere existence of data that is important, but our ability to make use of it. Experience has shown that as we make the associated metadata better organized and more complete, the more useful the underlying data becomes. In line with the “trustworthiness of data” and “ensuring the long-term sustainability…of high value data,” generous provisioning of metadata, including data provenance and data relations, is critical to enhance data sharing, to allow data to retain its usefulness over extended periods of time, and to allow traceability of results. There is an unmet need to better document workflows that create, transform, or disseminate data and to capture (and later present) data provenance, enabling scientists to answer the questions “who, what, when, how and why” for each data element; provide information about the connections and dependencies between the data elements; and allow human or automatic annotation for any data element.

2) In addition to “schema on read” (where data is applied to a plan or schema as it is pulled out of a stored location) data analysis and handling tools like Hadoop and MapReduce, NITRD should explicitly incorporate “schema on write” (where data is mapped to a plan or schema when it is written) metadata centric data analysis and handling tools in their VISION STATEMENT. Metadata centric “schema on write” techniques are critical to meeting NITRD goals of ensuring data consistency and trustworthiness along with being the only techniques capable of meeting NITRD member agency concerns such as data identification, access, reproducibility, provenance, curation, unique referencing, and future data availability.

3) Incorporate best practices in Big Data pioneered at NITRD agencies such as the DOE (SLAC BaBar) and NASA (HST and Kepler), that due to the nature of their work, have already encountered and overcame many of the Big Data problems that will be met more broadly by NITRD agencies and incorporated in the NITRD VISION STATEMENT.

4) Where possible, NITRD should encourage the use of existing, cost effective, supported, commercially available tools that meet NITRD member agency requirements and discourage expending NITRD agency resources on developing redundant internal tools that are difficult to sustain and leverage across the multiple NITRD agencies.

5) The ability to orchestrate global data intensive NITRD agency workflows and access and manage data on multiple storage devices, anywhere in the world is required. NITRD agency data needs to be easily and securely shared among globally distributed teams within NITRD member agencies. Data needs to move automatically to various workflow resources, based on policies, so data is always available at the right place, at the right time, and at the right cost — while keeping an audit trail as data is ingested, transformed, accessed, and archived through its complete lifecycle.

6) Like a needle in a haystack, high value data stored in NITRD agency storage systems can be effectively lost over time – stranding this data and losing its value forever. Metadata-based tagging and tracking of valuable information is needed so NITRD agency data can be found and analyzed even if it resides on very different, incompatible platforms anywhere in the world.

7) It’s essential to maintain NITRD member agency data provenance, audit, security, and access control in order to track data within workflows, through all transformations, analyses, and interpretations. NITRD agency data needs to be optimally managed, shared, and reused with verified provenance of the data and the conditions under which it was generated – so results are reproducible and analyzable for defects.
8) It’s imperative for NITRD member agencies to restrain storage costs by incorporating tools and procedures to: • Prevent worthless data from entering NITRD agency workflows and being stored • Migrate data to lower cost storage tiers using workflow policies where possible • Remove data that’s no longer valuable • Consolidate and automate complete data lifecycle management across multiple NITRD agency resources, avoiding costly individual resource management and administration

GTech

Srinivas Aluru, Professor Georgia Institute of Technology

Personal Information:

Srinivas Aluru, Professor College of Computing Georgia Institute of Technology 1336 Klaus Advanced Computing Building 266 Ferst Drive, Atlanta GA 30332. Ph: 404-385-1486 Email: aluru@cc.gatech.edu

Big Data Related Experience and Role in Big Data Innovation Ecosystem:

Georgia Tech (GT) is a top ranked public technology and science research institute. All of its computing and engineering departments are ranked in the top 10 by the U.S. News & World Report, with over half in top 5. We are the largest producer of engineering degrees awarded to women and underrepresented minorities. Research and education at GT is known for its real-world focus, and strong ties to industry. We are located in Atlanta, the eighth largest economy in the nation and a central hub for southeast industrial sectors. A large number of GT faculty have active research programs in big data, spanning applications and foundations, including collaborations with industry, government, and other Atlanta universities.

GT faculty projects have garnered support from all of the big data initiatives launched by major federal agencies (e.g. NSF Big Data, NIH BD2K, DARPA XDATA and ADAMS). I lead one of the eight NSF-NIH midscale big data awards made during the first round of big data federal funding. I have participated in the subsequent NITRD events including the first white house workshop on big data, and the Data to Knowledge to Action meeting. I co-chair the GT Faculty Council on Data Science and Engineering, which is charged with developing and executing a strategic plan for GT in big data. We are working towards a new Interdisciplinary Research Institute focused on data engineering and science, accompanied by an Innovation Hub for supporting industry engagement and entrepreneurship for translating GT big data research and expertise into economic and societal benefits. To support these activities, we are embarking on the construction of a 20 floor, 480,000 sq. ft. office tower augmented with 80,000 sq. ft. data center. The building will provide a long term base for GT efforts in high performance computing, big data, and computer modeling and simulation. The project will be a public-private partnership, with GT leasing about half the space and the rest made available for like-minded industry partners. Through this academia-industry co-location effort, we will drive economic development, creation of new jobs, technology engagement and transfer. We anticipate occupancy in 2017-2018.

Comments and Suggestions:

National Data Repositories

Access to data is one of the biggest impediments to big-data research. Many of the largest datasets originate in industry, government, or healthcare settings, and are not readily accessible due to issues such as privacy, loss of competitive
advantage, etc. This is rightly identified as a priority issue in the Big Data National R&D Initiative. A similar problem existed in the high performance community decades ago, which was effectively solved through national resources such as NSF TeraGrid, NSF XSEDE, shared DOE supercomputing facilities, etc. An effort of similar magnitude should be launched focused on data sharing, hosting community repositories, and supporting the compute-in-place paradigm. Where needed effective anonymization and identification/de-identification research should be supported to facilitate pervasive access to realistic datasets. Universities can act as nodal agencies to support such a national network. Georgia Tech is willing to be a partner in the area, and we are developing physical infrastructure capabilities to support such an endeavor.

Foundations on Big Data Analysis

Currently, much of the big data research is carried out in an adhoc manner, in the context of applications where data constitutes an overwhelming challenge. As the field progresses, unifying themes and common techniques with broad applicability are bound to emerge. Federal initiatives are already supporting research into core technologies. This should be further emphasized to encourage the fast development of broad foundations to support big-data research and development. Initiatives should target proposals that lead to development of paradigms accompanied by efforts to prove applicability in multiple domains. The analog of broadly applicable computability theory, and widely taught and used algorithmic paradigms, do not yet exist in the data domain.

Privacy and Ethical Considerations

Establishment of policies and guidelines for data sharing, along with policies for their responsible use, can remove the uncertainty in data sharing and accelerate big data research. Recent research and events such as the OSTP-MIT workshop on big data privacy explored privacy protection technologies, unsuspected breaches of privacy that can occur through seemingly de-identified big data, and rigorous computer science foundations for controlling access and ensuring privacy. Moving forward, it is important to establish uniform and legally acceptable protocols across all areas in the realm of big data, to remove impediments to research. At the same time, mandates for sharing big data where access to such data by the broadest possible community of researchers is vital for the public good, should be considered with appropriate safeguards. Specific examples that may fit this scenario include genomics driven medical care, and access to community health data that can predict impending outbreaks and environmental degradation that is responsible for illnesses.

HDF

Stable API's Are Critical to Evolving Data Storage Architectures

Dr. Ray E. Habermann, Director of Earth Science, The HDF Group, thabermann@hdfgroup.org

John Readey, Director Tools and Cloud Technology, The HDF Group, jreadey@hdfgroup.org

Dr. Habermann worked at NOAA’s National Geophysical Data Center for twenty--five years leading the development of data management and access systems for a wide--variety of NOAA observations. His experience spans CD-ROMS, scientific data formats and tools, relational and spatial databases, and web services. He is an active participant in development of ISO and Open Geospatial Consortium standards for geographic data. John Readey has worked at Intel
from 1997 through 2006 were he developed the Intel Array Visualizer – an application and library for array data visualization. From 2006 through 2014 he worked at Amazon where he developed service-based systems for eCommerce and AWS. They both recently joined The HDF Group to leverage their commitments to long-term data access and understanding.

The HDF Group (http://www.hdfgroup.org) developed and maintained the Hierarchical Data Format that has been used for scientific and engineering data and model results for more than 25 years. HDF is a foundation for standard formats serving scientific and engineering communities in a broad range of disciplines (Habermann et al., 2014). The HDF5 format supports data of any size, shape or source. It also supports the metadata that are critical for understanding the data currently and in the future. HDF is the primary format for all NASA Earth Observations. As netCDF4, it is the primary format for many climate model outputs and NOAA coastal, oceanic and atmospheric data sets. As Bathymetric-Attributed Grid (BAG), it is the standard format for NOAA Hydrographic Surveys.

Technical evolution always involves organizational as well as technical elements. A migration path that preserves existing capabilities during a transition is critical for facilitating progress on difficult organizational change associated with such evolution. The user communities mentioned above, and many others in the U.S. Federal workforce, have well-developed tool sets and workflows that are built around data in HDF5, the flagship HDF format. The creators and stewards of these high-value capabilities will resist evolution without a migration path. These people and systems form a significant obstacle to migrating the Federal environmental data infrastructure forward into the Big Data Paradigm.

Creating this migration path is critical to success in this challenging task, but migration path considerations are not included in the Visions and Priority Actions document.

The HDF Group has addressed the need for a migration path using an architectural concept termed the Virtual Object Layer (VOL) between the application I/O library and the underlying storage system. It controls how and where HDF5 data objects are actually stored. Analysis and visualization applications using the HDF5 library are shown in the upper left corner of Figure 1. These applications use the HDF5 library, available in many programming languages, to read and write data objects in local HDF5 files. These applications, and their existing API’s need to be reliable and stable through the transition to the new paradigm.

As data migrates into the cloud, across the bottom of Figure 1, the VOL provides this stability, allowing applications to read and write data and metadata in any underlying architecture. At the same time, new applications (e.g. Hadoop or other cloud tools) can access the cloud data using API’s appropriate for them.
The HDF5 format supports rich metadata associated with any data object in the file. The VOL approach writes/reads that metadata in a way that is consistent with the underlying storage strategy. For example, for a data repository hosted on Amazon AWS it would be useful to store metadata (Attributes, Link names, Types) in DynamoDB (a NoSQL database) while storing dataset values in S3 (an object storage system).

In conjunction with traditional HDF5 applications this approach offers new opportunities that may be effective in expanding access to large data collections: 1. An API can provide access to interactive web based apps for PCs/Tables/Smart Phones 2. NoSQL databases can be queried directly to effective search for specific attributes 3. Hadoop-based systems can import data directly from the object store 4. Systems such as Apache Solr can be leveraged to provide text based search

Habermann, Ted; Collette, Andrew; Vincena, Steve; Billings, Jay Jay; Gerring, Matt; Hinsen, Konrad; Benger, Werner; Maia, Filipe RNC; Byna, Suren; de Buyl, Pierre (2014): The Hierarchical Data Format (HDF): A Foundation for Sustainable Data and Software. figshare. http://dx.doi.org/10.6084/m9.figshare.1112485 Retrieved 00:42, Oct 15, 2014 (GMT)

HP

For further information please contact Jeff Graham at jeff.graham@hp.com or 703-896-0932.

Hewlett-Packard Company (HP) appreciates the opportunity to provide our point of view on the formation of a National Big Data Research & Development (R&D) Strategy. Invention and innovation are heritage values for HP. Today we’re investing more in R&D than ever before, as evidenced by our yearly investment of over $3 billion in R&D, and our portfolio of over 36,000 patents. Combining the research of HP Labs with the advanced development capabilities of our business units, we are rapidly commercializing our ideas.

HP’s view of big data—from an R&D perspective—sees a world facing a data explosion that is complex and is challenging to navigate and manage. The demand for computing is quickly outstripping the capability of many organizations’ underlying technology. HP’s focus on innovation is to bring advancements across the spectrum of business technology to tame the challenges of big data. This includes R&D in infrastructure, software and innovative
processes that effectively enables storage, processing, and security of all types of the information coming at us—structured, semi-structured and unstructured. Our approach is to breed new solutions for enterprises to gain an advantage by driving toward relevant information that is going to drive actionable intelligence.

HP considers the **transformation of computing** among the most high-impact ideas at the frontiers of big data R&D. It is vital to take the deluge of data and find the fastest, most secure and efficient route for gaining real value from that data. We are heavily committed to addressing this problem. One significant project—called The Machine—is a multi-year, multi-faceted research program that aims to fundamentally redesign computing to handle the enormous data flows of the future.

This initiative builds on research in which HP Labs already leads the world—from storing, managing and processing data, to being able to analyze and mine it to gain meaningful insights at the right time, for improved business outcomes. With one output called Moonshot, we’ve created system-on-a-chip packages that combine processors, memory, and connectivity. In our photonics research, we’re using light to connect hundreds of racks in a low-latency, 3D fabric. Our work in Memristors points to the development of universal memory—memory that collapses the memory/storage hierarchy by fusing the two functions into one hyper-efficient package. And we are breaking new ground in how massive volumes of diverse data can be analyzed, visualized, understood and converted into actionable intelligence by anyone in real-time. HP is building new software-defined systems that will fuse memory and storage, flatten complex data hierarchies, bring processing closer to the data, and enable security of systems at the point of attack. We’re also shaping the future of cognitive computing by making processing available on a massive scale to improve human decision-making.

The purpose of tackling big data technology is to advance the outcomes of everyday business processes, as well as confront the tough problems that make a real difference to citizens and the environment. These are enabling scientists to collaborate on big data for medical research helping Veterans to receive the best care in the most efficient manner, protecting citizens in emergency situations, and understanding the impact of environmental changes to impacting our wildlife ecosystem. HP’s philosophy is to create technology and services that “Make it Matter”.

HP considers The National Big Data R&D Initiative: Vision and Priority Actions draft document comprehensive, in that it contains key elements warranting serious consideration by NITRD agency leaders. HP concurs that streamlined access to relevant data sets and resources, development of partnerships that span private and public R&D sectors, actionable shaping of next-generation educational programs, and creation of gateways that enable the sharing of ideas and capabilities are indeed critical to enabling the Program’s vision.

We encourage the NITRD Program to consider incorporating several additional foundational elements to its national big data R&D framework, to include facilitating and nurturing communities of interest, the formation of accessible platforms, and sponsorship of big data R&D workforce retention initiatives.

- The concept of community has always been at the heart of research. We envision immense value can be achieved by fostering a national hub that facilitates and incentivizes the creation of communities of interest around federal big data R&D priorities. This initiative—leveraging knowledge management at its core—would serve to “match experts” and “match projects” in order to accelerate sharing and knowledge to realize further advancements. In fact, technology can be applied to assist in this effort and at HP we have piloted a capability to leverage contextual information being shared in the enterprise to identify expertise across our technologists to foster community and collaboration to solve problems. We suggest such an activity would be enabled by the creation of a national big data R&D registry that includes each project’s methodology, metrics and data-driven results.
• We believe research into new approaches, models and technology, and the application of those tools to federal mission challenges, could be facilitated by establishing common experimental platforms. There, multiple agencies could collaborate on exploring and testing new big data technologies and models without upfront capital investment and could be a springboard for enhanced collaboration. This could also reduce the “data” footprint by establishing a forum for large data sets and reduce data “churn”. Through supported platforms, we envision that the NITRD Program and private industry together could devise incentives for industry to contribute innovations and partner in the development of transformative solutions.

• Today, Data Scientists are in great demand and will continue to grow. We anticipate NITRD agencies are considering a variety of means within their authorities to retain such valuable staff, spanning compensation and benefits, continuing education, and unique mission experiences. NITRD should continue to encourage public programs and certifications for key roles for a data-driven organization. We envision the NITRD Program complementing agency initiatives by offering challenging and interesting projects, internal government competitions and prizes, and other opportunities for these employees to share successes and failures with motivated peers.

HP is an industry pioneer and global leader in the information technology market with a broad and deep portfolio of capabilities. We invest over $3 billion in R&D each year and have hundreds of dedicated researchers who are looking at emerging trends to understand where our world is headed. We are inspired by the opportunities because we know what the power of creative thinking and technology can do to transform lives, businesses, and communities. For further information please contact Jeff Graham at jeff.graham@hp.com or 703-896-0932.

Intertrust

Contact Information:

Dr. W. Knox Carey, Vice President, Technology Initiatives, Genecloud/Intertrust Technologies Corporation, 920 Stewart Drive, Sunnyvale, CA 94085. Telephone: (408) 616-1666. Email: knox@intertrust.com.

Our Role in the Big Data Ecosystem

The Genecloud project is an initiative of Intertrust Technologies Corporation, a private company in California focused on technologies for preserving individual privacy in the analysis of big data. The company is engaged in a number of technical initiatives in various application areas, including behavioral targeting for advertising, privacy for automotive data, home energy monitoring data, and others — areas in which analytics from multiple parties are essential for providing services and improving quality of life, but also areas in which compromise of personal data can cause great harm.

Healthcare information is arguably among the most sensitive personal information in this category. Our Genecloud project is focused on finding a pragmatic balance between access to healthcare data for analysis and patient privacy. Research in this project builds on work in trusted distributed computing, cryptography, policy management, information flow control, bioinformatics, and many other disciplines.

Comments on Visions and Priority Actions

We believe that the vision needs to be expanded to encompass the rights of individuals. Every datum in a Big Dataset was collected from a person whose rights must be respected — the right to remain anonymous, the right to negotiate the terms under which private data are used, the right to withdraw information from a data set. For reasons of expediency
these issues are often ignored or traded away, but technical solutions exist that allow the parties that rely on big data collected from individuals to solicit and enforce their consent for the intended uses.

These techniques are needed particularly when multiple agencies are involved in data analysis. While individuals may be willing to consent to certain uses of their private information by a first agency, this trust does not necessarily extend transitively to additional agencies, no matter how trusted by the first.

In addition, we believe that the second vision point (i.e. to understand trustworthiness of data and resulting knowledge) is essential. Analyses — and especially analyses that are made by a series of collaborating entities — should be accompanied by a digital chain of handling that allows all relying parties to understand precisely how, when, and by whom results were derived. Ideally, this manifest would also include code, initialization parameters, and other information that would aid in scientific reproducibility and post facto validation of results. It is a primary goal of our research to develop such tools.

In terms of research, education, and infrastructure priorities, we would note that very often issues of privacy and data security are treated as secondary priorities, to be addressed once an initial research goal is achieved. However, because security is a cross-cutting systemic concern, retrofitting data protection, trust management, and policy to an existing system seldom succeeds. The Internet itself is a case in point: the retroactive application of security into systems such as DNS, SMTP, and HTTP have been difficult and error-prone. To design more resilient systems, designs must incorporate specific security and privacy goals from the start, and this begins with education and infrastructure investments.

The federal government should lead by example on these issues. Private companies whose businesses depend on data collected from individuals have often neglected privacy — and in some cases acted explicitly against it — in their quest for larger data sets. In its inter-agency programs and in joint governmental/commercial projects, the federal government should insist that security and privacy controls be in place as a condition for funding or engaging in any collaboration. In addition, the federal government should set research and funding priorities to further these goals.

Privacy is a Fundamental Component of a Strategic Plan on Big Data

Big Data technologies will fundamentally change society, ushering in innovations that will improve our lives in ways that we can scarcely imagine. In the rush to adopt these technologies, however, we must be cognizant of their effects on individual privacy.

In healthcare, in particular, we must not be asked to choose between two equally fundamental rights: the right to lifesaving information and the right to personal privacy. The technologies needed to preserve and manage privacy do exist, and need not interfere with the analysis and sharing of big data. Indeed, we would argue that the presence of privacy guarantees will actually increase data sharing by mitigating many of the most serious risks.

We respectfully request that the National Big Data R&D Initiative incorporate research, educational programs, research goals, guidelines, and explicit requirements around individual privacy and policy control into its strategic plan. Our organization is committed to these goals and stands ready to assist in any capacity.
November 14, 2014

To: Wendy Wigen (wigen@mitrd.gov)

National Coordination Office for Networking and Information Technology
Research and Development National Science Foundation

Thank you for the opportunity to provide comments on “The National Big Data R&D Initiative: Vision and Priority Actions”. Having reviewed the initial framework, we have responded in the format requested below:

Respondent: Robert Detrick, President, Incorporated Research Institutions for Seismology

Contact: 1200 New York Ave. NW, suite 400, Washington, DC; detrick@iris.edu; 202 682-2220

Experience: IRIS manages data from hundreds of seismographic networks, collectively constituting one of the largest scientific archives of globally distributed observational data in the world. IRIS offers a wide and growing variety of services that Earth scientists rely on in over 150 countries worldwide, increasingly through web services that also deliver a wide range of data products for researchers and public outreach. IRIS manages instrumentation programs that play a key role in facilitating efficient collection of Earth science data and metadata.

What are the gaps that are not addressed in the Visions and Priority Actions document?

The absence of the U. S. Geological Survey or any Department of the Interior agency from the “NITRD Agencies” is a gap that risks lack of attention to a wide variety of in situ sensor data for the Earth sciences. The USGS and other DOI agencies work in cooperation with a variety of industries in the commercial sector, and closer cooperation could facilitate development and implementation of new concepts especially well suited to that type of data.

What do you think are the most high impact ideas at the frontiers of big data research and development?

The petroleum industry and other commercial groups have long dealt in very large volumes of Earth science data from ground-based sensor systems, such as 2-D seismic reflection data in the 1980s, and 3-D or 4-D seismic data today. After their proprietary economic value is gone, industry Earth science data are very often lost rather than re-used for public good, partly because of uncertainty about legal ownership. In addition, however, it would be costly to overcome technical hurdles to recover data that were usually not managed for long-term use, and thus were often left in formats and on media that became outdated.

In contrast, Earth science data collected and managed by the U.S. government generally have been managed for ongoing use. Earth science data that have long been important big data include a variety of satellite imagery; gravimetric, synthetic aperture radar, and other geodetic data; and other measurements from remote sensors. Free access to such data has been good for broader society, not only through immediate applications such as weather forecasting and mapping of natural resources but also through fundamental
Introduction

The MITRE Corporation is a not-for-profit company that runs Federally-Funded Research and Development Centers (FFRDCs) for the U.S. government. FFRDCs are unique organizations sponsored by a government agency to assist with research and development, studies and analysis, and/or systems engineering and integration. Formally established under Federal Acquisition Regulation (FAR) 35.017, FFRDCs meet special, long-term research and development needs that are integral to the mission of the sponsoring agency – work that existing in-house or contractor resources cannot fulfill as effectively. FFRDCs are required to operate in the public interest, be free from conflicts of interest and are operated, managed, and/or administered by universities, not-for-profit organizations, or industrial firms as separate operating units.

MITRE is unique in that our sole business is the operation of FFRDCs as well as in the breadth of FFRDC support we provide throughout the federal government. MITRE’s seven FFRDCs serve agencies in a variety of areas that impact the public in direct and indirect ways, such as national security; aviation safety and administration; tax administration; homeland security; healthcare; benefits services; cybersecurity partnerships; and other missions.
We are pleased to respond to your request for information regarding big data R&D based on our broad perspective gained from assisting numerous federal entities on big data issues, and from the unique perspective of a systems engineering company that combines a strong technical base with an informed awareness of the larger policy and context in which government operations are conducted.

**Vision Statement Thoughts**

While MITRE generally agrees with the thoughts conveyed in the draft vision statement, we are concerned that its length and depth will limit its utility. The best vision statements are concise, understandable to outsiders, inspirational to insiders, and clarify the path forward. The draft vision statement is an amalgamation of disparate ideas and platitudes that makes it difficult for readers to truly understand the administration’s goals for Big Data R&D. MITRE believes that the administration would be better served with a vision statement similar to the following:

> Our vision is a public-private innovation ecosystem that fosters collaboration on advancing Big Data technologies that can provide revolutionary capabilities and meaningful information for national priorities in a privacy-sensitive manner.

With a vision statement such as this, NITRD can ground the vision in practical terms using many of the statements found in the original draft and its succeeding bullets.

**Comprehensive Framework Thoughts**

Big Data research is taking place throughout the nation – in federal labs, in universities, and (mostly) in the private sector. Interest and investment in Big Data research exists. The predominant remaining needs, therefore, are to organize the research so that it is focused on priority national needs, facilitate collaboration, and encourage knowledge management.

Organize Research. The NSTC, via NITRD, can take a leadership role because it has the necessary breadth of focus combined with the ability to strategically direct federal research funds. We recommend that the first step be an assessment of current capability gaps, followed by a prioritization of those gaps into a published research strategy. Private-sector research aligned with the strategy can be identified and federal investments can help ensure that the highest-priority elements are being adequately addressed.

Facilitate Collaboration. NITRD is uniquely positioned to catalyze broader collaboration among groups of researchers and between researchers and end-users. This is likely best done through a combination of focused workshops and championing collaboration through identifying “use cases” for groups to collectively address. NIST-led efforts such as the NCCOE and NSTIC/IDESG can serve as examples.

Encourage Knowledge Management. Sharing insights and knowledge gained throughout the big data innovation ecosystem will be critical to enhancing the development of this field over the current state. The technical community’s typical peer-reviewed paper/poster process moves too slowly to be the predominant method for this field, however, as Big Data researchers require more rapid sharing of discoveries and capabilities gains. Researcher needs would be better served by NSTC/NITRD-enabled events, repositories, and other real-time knowledge management approaches.

**Research Needs**
MITRE has identified two areas of research need that we feel were not sufficiently addressed in the draft “Vision and Areas of Interest” statement:

- **Privacy.** While Big Data presents multiple opportunities it also creates privacy concerns, especially when previously separate data sets are combined for uses beyond the original purposes or when de-identified data is combined with information that renders it identifiable. Privacy protections need to be built into system designs from day one rather than considered a “to-be added later” feature. Research into how to best do this needs to be undertaken, and then implemented throughout all subsequent big data research projects.

- **Availability of Data Sets.** One of the largest costs for Big Data research is developing and maintaining data sets (static, real-time, and synthetic). Efforts to identify and make these available to a broad range of researchers, ideally organized along the identified priorities, would present a significant advancement for the big data innovation ecosystem. Commonly-available datasets, especially when combined with “use case” challenges, have been shown to advance collaboration and spur innovation. Ensuring privacy and IP protection of these data sets will be critical.

**NCDS**

This response to the NITRD RFI is submitted on behalf of the RENCI-led National Consortium for Data Science (NCDS, [http://data2discovery.org](http://data2discovery.org)), an innovative initiative created to address the challenges of big data that offers an informed perspective on the issues outlined in the RFI. As a public-private partnership of academia, corporate partners, nongovernmental organizations, and governmental agencies, NCDS exemplifies an approach that spans sectors, domains, and organizations. NCDS identifies big data opportunities and research needs, potential solutions, and works to implement those solutions for the benefit of society.

Creating regional centers devoted to data science and solving the challenges posed by big data is central to a national-level big data strategy as outlined in the National Big Data R&D Initiative vision draft. The strategy of applying public funding to support long-term national research objectives, in this case data science research, has already proved successful. NIH used this approach when it funded Clinical and Translational Science Award centers and the NSF employed it to fund its Supercomputer Centers. We believe a similar investment for big data research, education and outreach centers will generate an equivalent return by advancing national big data R&D goals.

In our vision, these centers will be developed guided by a five-pronged vision. First, the focus of the centers should cross agency boundaries. Second, the centers should be focal points for innovation that leverage the best public-private collaborations to address leading-edge big data challenges through fundamental and applied research. Third, the centers should serve as knowledge hubs that capture best practices, best technologies, and related knowledge for dissemination and reuse. Fourth, the centers should work closely with the educational community, from high schools to graduate and professional schools, to develop curricula and programs that promote workforce development and maintain national competitiveness. Fifth, the centers should be regionally-oriented to facilitate local impact and local interactions. Finally, the centers should be created in a way that facilitates long-term sustainability, and outreach efforts to inform the general public and policy makers about the potential and limitations of Big Data should be an integral requirement of all centers.

The challenges associated with big data span sectors, agencies, and disciplines. On the supply side, federal agencies and the programs they fund generate terabytes, if not petabytes, of data each day. Where possible the centers should work to remove stovepiped cyberinfrastructure and barriers to using and sharing data. For example, in the Earth sciences each relevant agency, (NASA, NOAA, and NSF) develops its own data management and dissemination infrastructure. Centers that can examine ways to leverage common capabilities and approaches across agencies are
essential. On the demand side, end users increasingly demand data synthesis that cuts across traditional disciplinary
divides. Similarly, decision-makers will need to rely on increasingly sophisticated tools that integrate multiple levels of
models and observational data from an array of sources not limited to one agency or program. To comprehensively
address these types of challenges requires multi-agency approaches.

A set of nationally organized, regional big data centers, possibly implemented as regional innovation hubs, can promote
coordinated research and development. Fundamental data science-related challenges require focused research and
development in order to support a forward-looking, national-level big data strategy as outlined in the vision document.
For example, there is an unexplored presumption that all data should be saved for all time. This assumption needs to be
examined much more thoroughly. Do all data need to be saved? How is an assessment done? Sensors and
cyberinfrastructure may be able to create and store data with 64-bit precision. However, from a data use standpoint, this
level of precision may not always be necessary. At a minimum, additional research should be supported in the areas of
networking, metadata, semantic knowledge creation, linked data, unstructured data models, middleware, and data
security. A set of coordinated regional big data centers/innovation hubs would draw upon regional strengths to advance
R&D while rapidly disseminating best approaches among them. Furthermore, these centers should support work that
transfers basic research into the applied domain and operational environments. These fundamental research areas can
help to address critical gaps and challenges faces by federal agencies as they seek to implement the various
presidential and OSTP open data and open access policies.

Big data challenges include managing data that are significant in complexity, size, and velocity, and federal agencies
may not have adequate resources to address these challenges. Data-intensive federal agencies, (e.g. Commerce,
NASA, NOAA, NSF, USGS) will require a much better understanding of data life cycle issues related to big data and
their core missions. While data growth is exponential, it is not clear agencies have the resources to save, let along
manage, all these data. Data lifecycle also impacts data provenance, which is key element to ensuring the
trustworthiness of data. Our envisioned big data centers should also identify and create best practices related to big
data.

A central goal for these centers will also be facilitating the development of data science curriculum and workforce
development programs. Numerous analyses show the demand for data-related jobs far outstrips the supply of qualified
applicants. A network of data science centers should work to fill this void by producing curricula to support practical,
applied, professional, and research education from high school through graduate school and professional training. Data
as a science is in a position similar to computer science in the 1950s. There is growing recognition that data science
should be treated as a specialization inside and outside of the academy.

Finally, these proposed centers should not operate on a traditional 1 – 3 year or piecemeal funding model. They should
be formulated as a national-level investment in the future of U.S. scientific and economic competitiveness and as an
essential element of a national security strategy. The initial levels of funding should be sufficient to ensure their long-
term viability.

The NCDS is an exemplar that addresses all the data science challenges outlined above. The NCDS has implemented
innovative educational efforts and a Data Fellow program. It is establishing a first-of-its-kind Data Observatory and
Laboratory. The consortium marshals regional and national- level participants from both the public and private sectors to
address these challenges. As NITRD’s big data initiative moves forward, the NCDS welcomes the opportunity to serve a
model example, to share lessons learned, and to operate as one of the inaugural big data centers.
NERSC-SDF

Big Data Facility to Enable Data-Driven Science  Shane Canon, Prabhat, Katie Antypas, Jeff Broughton, Sudip Dosanjh NERSC, Lawrence Berkeley National Lab.

**NERSC’s role in Big Data:** NERSC has a proven track record of running a production quality national user facility for over 40 years. We provide high performance computing resources and user support to thousands of scientists spanning hundreds of scientific projects. Our users range from Nobel Prize winners to distinguished scientists to emerging early career researchers. For more than a decade we have engaged with big data communities in High Energy Physics, Climate and Genomics in supporting their data-centric workloads. We import experimental and observational datasets from telescopes, light sources, particle physics detectors and genome sequencers, and support complex data workflows that handle data ingest, pre-processing, analytics and visualization phases, resulting in scientific insight.

**Suggestions for NITRD Vision and Priorities Document:** We share the broad goals outlined in the NITRD document. We would like to emphasize that sustained investment in production facilities, reusable and cross-domain tools and infrastructure, data representation and storage mechanisms will be critical for long-term impact in the Big Data space.

**Proposed High Impact Ideas:** The Department of Energy is a pioneer in building and operating national users facilities and applying advanced computing to enable scientific discovery. This is illustrated by its premier computing facilities such as NERSC. To date, these facilities have been primarily focused on enabling large-scale modeling and simulation and access has been generally limited to projects linked to the DOE mission. With the increasing importance of data in enabling scientific discovery and driving scientific leadership, the nation requires a comprehensive and coordinated initiative to expand these national capabilities to address data-intensive scientific challenges. We envision this initiative as an integrated, multi-agency effort that couples cutting edge cyber infrastructure, next generation software services, and intensive training and consulting to enable scientists to address problems of national importance and maintain leadership in science.

**Facilities:** Scientists are increasingly limited in the problems that they can tackle due to limits in storage, analysis resources, and networking connectivity. The available resources are often mismatched both in design and scale to the challenges at hand. We envision a limited set of federated resources, potentially multi-agency in design, that are architected and optimized for the needs of the scientific community. Access to these resources will remove many of the barriers that exists today. Centralizing these resources will provide both efficiencies in operation, access to unprecedented scale, and enable a level of data integration that can not be achieved to date.

**Services:** A set of scalable, API-driven software services built on open standards is required to enable the hardware, ensure open-access, and foster collaboration across agencies. Investments are needed at all levels of the Data stack: from workflow and batch systems to co-ordinate data movement and job execution, to runtime analytics frameworks (e.g. BDAS, Hadoop), to user-facing, productive libraries (e.g. MLBase, GraphX). Finally, the ability to upload, host and share data nationally, and internationally will be critical; continued investment in portal and gateway technologies is required.
Expertise and Training: To enable scientists to effectively utilize these capabilities, they require increased access to experts and training. These training efforts should integrate with existing academic programs and leverage emerging online learning. Existing resource operators already possess significant skill in both data science and domain science and must be a part of the strategy. This includes developing and disseminating training material, but also providing more advanced consulting for critical challenges.

Recommendations on fostering new cross-domain and cross-sector partnerships: NITRD has done a commendable job of organizing various inter-agency working groups (Big Data, LSN, JET, MAGIC). Such working groups are a critical component in terms of gathering strategic input from a variety of perspectives and coordinating activities. We would recommend arranging in-depth town-hall meetings, or requirement gathering workshops which conduct deep-dive sessions into leading Big Data science use cases from various agencies, and explicitly target common research and production infrastructure related issues. Currently, there is significant redundancy in our collective approach towards exploring the Big Data space, and NITRD can play a leading role in proposing a common vision.

Relevance of NERSC to NITRD strategic plan: NERSC has over 40 years of demonstrated production experience in the High Performance Computing arena. Over 5000 users utilize NERSC facilities to produce breakthrough science results each year. While our primary scientific engagements stem from the Department of Energy, recent collaborations with NIH and NSF have positioned us to respond well to a broader class of science use cases and inter-agency collaborations. We believe strongly that we have the right people, skills and infrastructure to both provide input to NITRD strategic plan, as well as execute a Big Data vision at the national scale.

Niemann

Data Science for the National Big Data R&D Initiative Submission by Dr. Brand L. Niemann, Director and Senior Data Scientists/Data Journalist, Semantic Community, and Founder and Co-organizer of the Federal Big Data Working Group Meetup, October 10, 2014

Submission by Dr. Brand L. Niemann, Director and Senior Data Scientists/Data Journalist, Semantic Community, and Founder and Co-organizer of the Federal Big Data Working Group Meetup, October 10, 2014

• Who you are—name, credentials, organization: Dr. Brand Niemann, Director and Senior Data Scientist/Data Journalist, Semantic Community and Founder and Co-organizer of the Federal Big Data Working Group Meetup
• Your contact information: bniemann@cox.net
• Your experience working with big data and your role in the big data innovation ecosystem: Former Senior Enterprise Architect & Data Scientist with the US EPA, works as a data scientist, produces data science products, and publishes data stories for Semantic Community, AOL Government, & Data Science & Data Visualization DC Founded and co-organizes the Federal Big Data Working Group Meetup.
• Comments and suggestions based on reading the initial framework, The National Big Data R&D Initiative: Vision and Priority Actions, and guided by the following questions:
  ◦ What are the gaps that are not addressed in the Visions and Priority Actions document? A community of practice with a Mission Statement like the Federal Big Data Working Group has and that implements our version of the NSF Strategic Plan: Data Scientists, Data Infrastructure, and Data Publications using Semantic Web - Linked Open Data like the NITRD espouses.
From an interagency perspective, what do you think are the most high impact ideas at the frontiers of big data research and development? The Data FAIRPort, NIH Data Commons, and Semantic Community Sandbox for Data Science Publications in Data Browsers.

What new research, education, and/or infrastructure investments do you think will be game-changing for the big data innovation ecosystem? Data Mining all the research, education, and infrastructure investments for what has already been/is being paid for, especially for the agencies that have $100M or more in R&D funding per the so-called OSTP Holdren Memo, like the Federal Big Data Working Group Meetup is doing.

How can the federal government most effectively enable new partnerships, particularly those that cross sectors or domains? By participating in Communities of Practice on Data Science, Data Visualizations, Big Data, etc.

- A short explanation of why you feel your contribution/ideas should be included in the strategic plan: The Federal Big Data Working Group is what the Vision and Actions to be Taken say - a new partnership, a new gateway, and a long term sustainable activity.

NSIDC

Dr. Mark Serreze, Director NSIDC National Snow and Ice Data Center Cooperative Institute for Research in Environmental Science University of Colorado at Boulder 449 UCB Boulder, CO 80309

Our Experience Working with Big Data and Our Role in the Big Data Innovation Ecosystem

For nearly 40 years, the National Snow and Ice Data Center (NSIDC) has been at the forefront of cryospheric data management. We make data and information about the cryosphere and Polar Regions available and useful to researchers, policymakers, the media, industry, and the public at the behest of many agencies, including NASA, NOAA, NSF, and USAID. As leaders in data management, we are well versed in making data available across complex domains and disciplines, including remote sensing, traditional and local knowledge, and the social sciences. As a data center, we have contributed to education and training in data management at all levels through a number of activities, including the Federation of Earth Science Information Partners data management short course for scientists. We are also committed to the education of the next generation of data curators, through classes at the Graduate School of Library and Information Systems at UIUC and through the inculcation of undergraduate, graduate, and early career scientists from a variety of fields and programs into NSIDC activities.

For millions of people, NSIDC is the trusted source of information about Arctic sea ice and the Polar Regions. This trust was gained through our efforts to both provide thoroughly documented data and valueadded higher-level data and information products and tools, all accessible to other disciplines and communities. Because our staff includes domain scientists, who are noted researchers in their own right, a strong informatics/data management research team, talented technical writers, and software developers, we can provide a number of valuable services, including expert analyses of the data for multiple audiences, with a superb user services office to help people understand how to use the data. As such, NSIDC and other domain repositories fill a critical role within the overall data ecosystem.

Our Comments and Suggestions:
What are the gaps that are not addressed in the Visions and Priority Actions document?

- Recognize that a Big Data innovation ecosystem will be built upon existing resources, such as domain repositories and other initiatives, and that a primary initial effort should be to identify these resources and determine how to integrate them into the emerging ecosystem.

- Understand that there is a major gap between how business and science see Big Data. The language and definitions used by these communities varies. Additionally, the scope and types of questions from each of these perspectives diverge greatly; however, each community has much to learn from each other.

From an interagency perspective, what do you think are the most high impact ideas at the frontiers of big data research and development?

- Provide incentives and funding that allow for full documentation and metadata with the goal of increasing attention to data usability. This is a high impact area that is currently underfunded/undervalued. Our experience has shown us that without attention to data reuse, none of the NITRD values are obtainable. We advise making documentation, metadata, and data usability a funded part of all future data generation.

What new research, education, and/or infrastructure investments do you think will be game-changing for the big data innovation ecosystem?

- Support funding models for domain repositories that allows cross-agency and cross-sector data stewardship activities over the long-term rather than focusing on maintaining short term non-data stewardship-focused grants.

- Create cloud-based test beds for R&D with sample datasets, community supported algorithm libraries, with cross-agency access and support.

- Integrate the library and information science communities with the existing suite of domain repositories so that the experiences gained over decades within repositories becomes visible and is used within those communities.

- Establish rewards throughout the research enterprise for researchers who engage in good data management practices.

How can the federal government most effectively enable new partnerships, particularly those that cross sectors or domains?

- Facilitate activities that build upon the missions of multiple agencies to ensure better cross-agency funding coordination for data activities.

- Develop funding models for infrastructure and data management projects that provide funding for extended periods of time.

- Continue funding planning workshops and Research Coordination Networks to ensure that communities have the opportunity to connect and build collaborations around Big Data.

- Create new fellowships and programs for graduate students to enable both scientific research and the development of technical/data systems (perhaps something similar to the IGERT fellowships that produced researchers fluent in interdisciplinary research).

Why do you feel your contribution/ideas should be included in the strategic plan?

NSIDC covers the gamut of types of Big Data - from large volume satellite and airborne sensors (the MODIS L2 data set includes more than 7 million data, metadata, and browse files), to extremely diverse long-tail data sets (the ACADIS system has more than 2800 data sets), to observations made by Indigenous Arctic residents. Our data are extremely diverse, both qualitative and quantitative, and most of the data are openly available except where ethical or legal considerations prohibit it. We have experience tying widely disparate types of data and information together, including experience working with and connecting Local and Traditional Knowledge (LTK) and community data. While these data
are not as large as current standards, they are diverse ranging from instrument-based point samples, maps, text narratives, photographs, video, audio narratives and knowledge models. Increasingly, these data are being linked to other data, such as remote sensing and sensor measurements, held by NSIDC and other organizations. Additionally, we handle the whole range of data from the lowest level raw data through the development of higher level data products, such as interpretive products useful to the public and policy makers.

Examples, where appropriate

IceBridge example: IceBridge collects data from airborne sensors, ties it with ICESat, Cryosat, ICESat-2 big data, but also with data collected from diverse measurements, the Navy is making on the ground in selected locations.

Nugent

Peter Nugent, Division Deputy for Scientific Engagement Computational Research Division, LBNL Adjunct Professor of Astronomy, UC Berkeley M.S. 50B-4206 - 1 Cyclotron Road - Berkeley, CA, 94720-8139 Phone:(510) 486-6942 - Fax:(510) 486-4300 - Cell:(925) 451-4001 E-mail: penugent@LBL.gov - Web: http://www.lbl.gov/~nugent

Role and point of view in the big data innovation:

I began my research life as a theorist and have since transitioned to observational astrophysics. All of the major work I have carried out in my career has involved high-performance computing (HPC). I have been able to follow a variety of scientific fields, and the effect that changes in HPC architectures and tools has had on them over the past two decades due to my joint position at the National Energy Research Scientific Computing Center (NERSC) and at the Computation Research Division at LBL. Since 2008 I have focused almost exclusively on challenges in data analysis, specifically mining extremely large astrophysical data sets and confronting these directly with even greater sized simulation data. This RFI offers the opportunity to create a community from which one could leverage help to advance three major themes I see as crucial in the coming decade for data analysis: taking full advantage of new HPC resources (multi-core/many-core, BurstBuffer, high-speed networking, etc.); novel techniques for confronting experimental data with simulations; and devising courses at both the undergraduate and graduate level to educate our students on the best ways to address these problems in the coming decade.

Extreme Data Analysis in Cosmology:

In recent years astrophysics has undergone a renaissance, transforming from a data-starved to a data-driven science. A new generation of experiments — including Planck, BOSS, DES, DESI, Euclid, WFIRST and LSST — will gather massive data sets that will provide more than an order of magnitude improvement in our understanding of cosmology and the evolution of the universe. Their analysis requires leading-edge high performance computing resources and novel techniques to handle the multiple PB’s of data generated throughout these surveys. Furthermore, interpreting these observations is impossible without a modeling and simulation effort that will generate orders of magnitude more “simulation” data — used to directly understand and constrain systematic uncertainties in these experiments and to determine the covariance matrix of the data. As we enter this era of precision cosmology a thorough propagation of errors on measurements in both the experiments and simulations becomes essential. This is especially true in modern cosmological models where many parameters and measurements are partially degenerate, and such degeneracies can lead to important shifts in the cosmological parameters one is trying to measure.
Data Co-Design is a term that refers to a computer system design process where scientific problem requirements influence architecture design and technology, and constraints inform formulation and design of algorithms and software. To ensure that future architectures are well-suited for data target applications and that major data scientific problems can take advantage of the emerging computer architectures, major ongoing research and development centers of computational science need to be formally engaged in the hardware, software, numerical methods, algorithms, and applications co-design process. An example of this is engaging next-generation HPC centers in their engineering efforts on future big-iron machines they will deploy for use by the general scientific community. For example, the software development effort for BurstBuffer (a tier of solid-state storage systems to absorb application I/O requests) in order to more readily handle the analysis of ~PB datasets in cosmology as outlined above. As there will be many science applications that could benefit from such an effort, NITRD could aid in the expansion of the co-design process to more fields across several agencies.

**Simulation-Based inference for cosmology:**

The DOE, NASA and NSF have invested significantly in large-scale Dark Energy experiments (the Dark Energy Survey – DES, the Large Synoptic Survey Telescope – LSST and the Wide Field infrared Telescope - WFIRST) where observations of Type Ia supernovae (SNe Ia) play a central role. They have also been a driving force behind the computational science of SN Ia physics as well as large-scale simulations of the structure of the universe. While large-scale simulations have a long history of using simulations directly in the analysis of observational data, the SN field does not. Today, SN cosmologists still rely on fully empirical techniques to chase Dark Energy. These methods are derived from data that suffer important selection biases and calibration issues, using parameterizations that may be only qualitatively justified after the fact. The gulf between simulation and observation is especially disappointing given that many SN Ia systematic errors affecting Dark Energy measurements should be best (or may only be) addressed through simulation. This is an area that many experimental and observational programs lack as well. An effort to promote the infrastructure to bridge the gap between the simulation and modeling effort with data analysis can pay huge benefits.

Furthermore, we need to take a serious look at how these agencies share these large data sets and analysis tools. Currently it is everyone for themselves and whatever home-brew method they come up with. This creates a lot of hurdles in the joint analysis of data sets that span these agencies.

**Education:**

In the spring of 2013 I taught a course at Cal entitled, High Performance Computing for Astrophysicists. The goal of the course was to provide an introduction to Unix and the working environment on HPC systems. Students were given accounts at NERSC in order to gain experience running a variety of current parallel codes in astrophysics and handling both the resultant large datasets generated by these simulations as well as observational datasets through NERSC's Science Gateway Nodes. I think the time is ripe to work on creating a more standard set of courses in our university system, applicable to a variety of scientific domains, in which one can expose students to a variety of techniques in data analysis that highlight the power of novel algorithms, HPC architectures and the latest in I/O, databases, etc. Getting NITRD to promote this educational effort across the various agencies would create huge benefits in both the near and long-term future for data science.

https://semanticommunity.info/Data_Science/Data_Science_for_the_National_Big_Data_R_and_D_Initiative

Updated: Sun, 14 Jul 2019 10:27:14 GMT

Powered by mindtouch™
Our project ("Chemistry and Microphysics of Small Particles") is focused on the development and application of unique methods for single particle multidimensional characterization, to yield quantitative, comprehensive, real-time information on the properties and transformation of small particles that are ubiquitous in natural and human-made environments. These methods are applied to study particles of interest to basic and applied sciences, including nanoscience, nanotoxicology, mesoscience, atmospheric science, and combustion research.

The behavior and impact of small particles depends on a multitude of their properties, many of which are strongly coupled. The size, internal composition, phase, density, shape, morphology, optical properties, hygroscopicity, activity as cloud warm and ice nuclei, and others - all play a role. Unlike traditional approaches that rely on parallel measurements conducted on heterogeneous mixtures of particles, we simultaneously measure all relevant properties of millions of individual particles in real-time.

By necessity this multidimensional single particle characterization routinely produces vast amounts of high dimensionality data, on millions of particles, the classification, visualization, mining, and analysis of which calls for unconventional methods that must draw on statistical methods, while preserving the wealth and depth of information. Moreover, analysis should be based on data generated by all relevant instruments, and include the relationships between them, their temporal evolution, and, for data acquired on aircraft, a way to visualize it all in a geo-spatial context.

The challenges we face are clearly common to other fields that generate massive multidimensional, complex datasets that require matching advances in the science of data organization, visualization, and mining. To that end, we have been developing and applying novel approaches for analysis of large multidimensional complex datasets, in collaboration with Prof. Klaus Mueller (State University of New York at Stony Brook), a specialist in data mining and visualization.

Our unique data visualization and mining program, SpectraMiner, makes it possible to handle data on millions of particles organized into hundreds of clusters, limiting loss of information and thus overcoming the boundaries set by traditional data cluster analysis approaches. SpectraMiner organizes the data and generates an interactive circular hierarchical tree, or dendrogram, providing the user with a visually driven, intuitive interface to easily access and mine the data of millions of particles in real-time.

In addition, we developed ClusterSculptor, our expert-driven visual classification software, which uses a novel, expert-steered data classification approach to provide an intuitive visual framework to aid in data clustering. It overcomes the limitation of traditional statistical approaches by offering the scientist the ability to insert their expert knowledge into the data classification.

To visualize and analyze the relationships between multitude of observables and to do so in a geo-spatial context, we recently developed the interactive visual analytics software package, ND-Scope, that is designed to explore and visualize the vast amount of complex, multidimensional data acquired by our single particle mass spectrometers, along with other relevant datasets.
with all other aerosol and cloud characterization instruments on-board the aircraft. We demonstrate that the interactive and fully coupled Parallel Coordinates and Google Earth displays of ND-Scope make it possible to visualize the relationships between different observables and to view the data in a geo-spatial context.

Analysis of large data includes data pre-processing (data preparation, integration, reduction, and clustering) that often consumes, as much as 80-90% of the effort. Since data analysis can often be mission critical, pre-processing should be done effectively and fast. Recently we have developed a GPU-accelerated incremental correlation clustering of large data with visual feedback 6 that makes it possible to achieve significant speed-ups over the sequential clustering algorithms. It can be used to detect and eliminate redundant data points, offering a viable means for data reduction and data sampling, and find the outliers or a few “golden nuggets” in vast amounts of data.

Moreover, big data pre-processing, is rarely interactive, which stands in conflict with expert-users, who seek immediate feedback and answers. Our approach provides streaming visual feedback of the clustering process using Multi-Dimensional Scaling dynamic display, allowing interactive visualization and input by the user to fine-tune the clustering parameters and process. Note that visualizing results is also more intuitive than raw text or tabulated data and can increase the chance of spotting patterns that might otherwise go unnoticed.

All software packages we are developing have applications far beyond single particle characterization and are suitable for other large and complex, multidimensional data. While the pre-processing of large data may require the multiple-GPU approach or use of supercomputers, it is highly beneficial to perform visualization and mining of the data in interactive and intuitive manner through visual interface on personal computer. This approach is central to all our software packages, as it puts the user at the center of information flow and decision making.


PlanetOS

Contact Information
Company Background

Planet OS is a data discovery engine for sensor and machine data. Our mission is to index the real world data on oceans, land, air and space coming from sensors and robotic devices. Offshore oil & gas companies and governments are able to implement Planet OS to index, discover and make sense of their growing sensor data using a hybrid cloud and onpremise approach. Planet OS contextualizes, visualizes, analyzes and fuses datasets, making the data universally accessible with external APIs.

Comments and Suggestions

What are the gaps that are not addressed in the National Big Data R&D Initiative Visions and Priority Actions document.

The Initiative should address societal and business needs next to policy needs. Unprecedented access to data is transforming individual and collective privacy concerns, opening up completely new venues for businesses. International agreements in crosscutting concerns like individual privacy and use of data for profit, will ensure prosperity in a global competitive landscape.

From an interagency perspective, what do you think are the most high impact ideas at the frontiers of big data research and development?

Despite these technological advances and an accelerating trend in data volume, vast repositories of data are not readily accessible. Finding, formatting, aggregating and processing diverse types of information remains a significant challenge, in particular if the required data is distributed across multiple agencies. Federal Agencies have an opportunity to address this deficiency by utilizing Interagency Data as a Service, Software as a Service and Platform as a Service concepts. These service formulations are based on the idea that the product — data, software applications, or computing platforms — can be provided to the user ondemand, regardless of geographic or organizational separation of provider and consumer, and without regard to the physical system on which the data, software, or computing reside.

What new research, education, and/or infrastructure investments do you think will be gamechanging for the big data innovation ecosystem?

The key to big data innovation on the federal level is effectively indexing and distributing data for public and commercial use. Once the data is universally accessible, the scientific and private sectors can start creating added value by leveraging the cumulative value of multidomain federal data that has not been reasonably accessible so far.

Accordingly the most critical component to foster innovation is making federal data products available at scale and in realtime, so the investments should focus on research and infrastructure that enables this. The highest goal would be an open platform that involves scalable cloud infrastructure, powerful programmable interfaces, interactive tools that
streamline user experiences, automated data flows and an advanced database that complements federal core data acquisition systems.

How can the federal government most effectively enable new partnerships, particularly those that cross sectors or domains?

The government can ensure effectiveness by enabling private sector to develop and operate open data platforms, while keeping the data stewardship role within the agencies. As a strategic infrastructure, a big data grid should be operated by multiple vendors who are empowered and incentivized to ensure open access to infrastructure and data to third parties independently from the government. Government involvement should be focused on stewardship activities.

A short explanation of why you feel your contribution/ideas should be included in the strategic plan.

The Planet OS Global Open Data Initiative, Marinexplore.org, is a growing global partner network of 33 institutions offering fast access to a rich directory of wellorganized environmental data from 43,000+ data streams. Planet OS as a company has amassed unique experience in developing and operating a global data initiative since July 2012. The company’s experience in successfully operating the initiative with more than 7,500 participants is the foundation for the proposals put forth in this response to the RFI.

RedHat

1. Jon Shallant, Account Executive NSF, Red Hat
2. jshallant@redhat.com, (703)-629-0564
4. I believes the data life cycle in the enterprise is changing rapidly and requires an open, agile approach to innovation in the data center.
5. None at this time
6. Customers are looking for choices in open software solutions for big data and hybrid cloud to help them easily and quickly transform their infrastructure and applications.
7. Hadoop and OpenStack are key components to big data.

- **Cloud-ready data platforms** to accelerate the transition to the open hybrid cloud with Red Hat Enterprise Linux OpenStack Platform and Sahara integrated with Cloudera Director and Cloudera Enterprise, all managed by Red Hat CloudForms.
- **Enterprise-ready data platforms** that are secure, scalable and easy to manage with the integration of Red Hat Enterprise Linux, OpenJDK support and Red Hat Storage Server with Cloudera Enterprise, Cloudera Manager and Cloudera Navigator.
- **Agile data integration and application development tools** with Red Hat JBoss Middleware and OpenShift by Red Hat integrated with Cloudera Enterprise that leverages the Cloudera Kite libraries, Cloudera Impala and Apache Hive connectors.
8. Investing in the Open Source community/contributors. The Open Source community is made up of businesses, government, academia, and individuals. Example, Red Hat, Cloudera and Intel recently partnered in order to give customers an open and modular technology stack to quickly derive new insights from their data, optimize their existing investment in platform infrastructure and lower the overall cost of managing data platforms. The Open Source community allows more collaboration to help innovate and lower costs of technology. It also solves a business need rater than building software around non essential features.

9. Cloudera, the leader in enterprise analytic data management powered by Apache Hadoop®, and Red Hat, Inc. (NYSE:RHT), the world’s leading provider of open source solutions, announced an alliance to deliver joint enterprise software solutions including data integration and application development tools, and data platforms. By integrating a broad range of products and technologies, Cloudera and Red Hat will help customers harness the fast changing big data life cycle with open, secure and agile solutions.

10. Case Study OpenStack:
   - Cloudera Case Studies: [http://www.cloudera.com/content/en/r...ase-studies&q=](http://www.cloudera.com/content/en/r...ase-studies&q=)

RTI

Paul P. Biemer, Alan R. Blatecky, Alan F. Karr (RTI International)

**Introduction.** Data are everywhere. They being produced by virtually all scientific, educational, governmental, societal and commerce organizations, and being generated by surveys, mobile and embedded systems, sensors, observing systems, scientific instruments, publications, experiments, simulations, evaluations and analyses. In this response, we highlight three central Big Data issues—curation, data quality, and privacy/confidentiality, especially the need to balance confidentiality with usage.

**Experience.** RTI International is a world leader in the collection, integration, analysis, interpretation and visualization of complex data, in contexts ranging from longitudinal surveys to physical measurements to social media to program evaluation, and beyond. The Institute's interests and activities span the full pathway from data to information to knowledge to decisions and practice.

**Curation.** Data that are useful for analysis must be available in ways that allow them to be searched, analyzed, and manipulated, across disciplines, domains, and geographic boundaries. Unless data are adequately curated, it will be very difficult for science, government or business, to use—let alone re-use—them. Federal agencies spend a considerable amount of funding and efforts generating data, but much less time and effort ensuring that the data will be useful to others. Data that are not adequately described with metadata will be very difficult to locate and interpret. Data not identified by permanent identifiers such as an OID (Object Identifier) cannot be effectively registered. Absent information on how the data have been modified or altered over time (addressing issues of provenance and versioning), problems of reproducibility may be insurmountable.
The point is this: if data are not adequately curated, it is questionable whether they should have been generated in the first place, as it will be very difficult to use them in the future or in conjunction with other data. Since a major assumption of the National Big Data R&D Initiative is re-use of all types data, it is strongly recommended that Federal agencies specifically, and more forcefully, address data curation issues (e.g.,metadata, provenance, versioning, permanent identifiers) in their Big Data programs, projects and activities.

**Data Quality.** Some of the errors that plague Big Data are well-known. As they are created, Big Data are often selective, incomplete and erroneous. New errors can be introduced downstream as the data are cleaned, integrated, transformed, and analyzed. Data munging (wrangling) steps comprise 50-80% of the work involved in getting a dataset ready for analysis, but these steps can add both variable and systematic errors to the data, resulting in unreliability, invalidity, and biased or even incorrect inference. To illustrate, Big Data are often portrayed as allowing detailed examination of extremely rare phenomena. But, even very small false identification rates (on the order of one hundredth of one percent or less) can contaminate these data, rendering them essentially useless for analytical purposes. Even in the absence of errors, Big Data pose new challenges to inference, such as noise accumulation, spurious correlations, and incidental endogeneity (Fan, Han, and Liu, Challenges of Big Data Analysis, National Science Review, doi:10.1093/nsr/nwt032, 2014). Data errors can exacerbate these problems.

To illustrate, in high-dimensional data, spurious correlations are not rare: in simulated 800-deminsional data having no correlations and relatively small sample sizes, the analyst has a 50% chance of observing an absolute correlation that exceeds 0.4. In the presence of data errors, these risks increase substantially even for sample sizes in the hundreds of thousands. Systematic errors common to two variables magnify the observed correlations (Biemer and Trewin, A Review of Measurement Error Effects on the Analysis of Survey Data. In L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwarz, and D. Trewin (eds), Survey Measurement and Process Quality, New York: Wiley and Sons, 603-632, 1997), thereby exacerbating the problem. It is incumbent upon the Federal government as a major user and supplier of Big Data to be cognizant of the issues associated with Big Data errors and their consequences. Failure to recognize the limitations of errors in these data will lead to false discoveries, failed policies and errant decisions.

Balancing Privacy, Confidentiality and Data Availability. For many years, official statistics agencies such as the Census Bureau have struggled to balance legal and ethical obligations to protect confidentiality of their datasets, yet make data available for research and policy purposes. Today, several developments are converging that make the balancing increasingly challenging. First is the scale of the data: the more that is collected, the easier it is to identify records, especially by linking to other datasets. Second is the increasing extent to which data subjects are unaware that data about them are being collected, to the point that extant foundations of informed consent being questioned. Repurposing of data creates further issues: protection in one context does not imply protection in all contexts. Finally, “big computation” threatens all existing methods of disclosure limitation, by casting disclosure risk as a problem of computational complexity—not whether confidentiality can be broken, but how much computational effort is needed to do so.

Despite changing attitudes regarding privacy, not protecting data will not become an acceptable, or even the prevailing solution. Indeed, in light of near-daily reports of breaches in the security of corporate information systems, there may a backlash. For multiple reasons, every government data collection in every country in the world is facing severe problems with declining response rates. Not just new techniques, but even new abstractions of disclosure risk, disclosure harm and data utility are needed.

https://semanticommunity.info/Data_Science/Data_Science_for_the_National_Big_Data_R_and_D_Initiative

Updated: Sun, 14 Jul 2019 10:27:14 GMT

Powered by mindtouch™
Adding to these pressures are recent mandates (some unfunded) to make research data available, especially given the seemingly excessive rate of “failure to replicate.” Failure to pay proper attention to curation, data quality and privacy and confidentiality will attenuate or even negate entirely the benefits of making data available. It is all too easy to make data available but unusable.

**Conclusion.** It is important to realize that data are merely the first step—albeit an essential one—in the pathway from data to information to knowledge to decisions. The value of data is not inherent, but rather is derived from the inferences, insights and decisions drawn from the data. Today, principled ways of trading off cost and data quality are only emerging; the Federal government can play a leadership role by moving attention to decision quality. More data are not necessarily better, if they are not properly curated, their quality is not understood, or they are not protected: bigger is not necessarily better!

Nor do Big Data reduce, let alone eliminate, the crucial need to characterize and quantify uncertainties in inferences. Indeed, as noted above, Big Data can make the need more urgent. There are serious risks of a society drowning in data and starved for principled research and sound policy derived from the data. Federal government initiatives may be the only way to prevent this from happening.

**SDSC**

**Submitted by:** San Diego Supercomputer Center (SDSC), University of California, San Diego, 9500 Gilman Dr., #0505, La Jolla, CA 92093  **Contact:** Michael Norman, PhD, Center Director Tel (858) 822–5450, email mlnorman@sdsc.edu

**Experience working with Big Data and role in Big Data innovation ecosystem:** SDSC is a major academic supercomputing center funded by the NSF with a long history in Big Data innovation, shared community infrastructure and impactful, cross domain, Big Data projects ranging from WIFIRE to OpenTopo and the Neurosciences Information Framework (NIF), all underpinned by Big Data resources such as Gordon and soon, Comet.

Comments and Suggestions On The Initial Framework

What are the gaps that are not addressed in the Visions and Priority Actions document?

A strategic plan should recognize that Big Data systems are fundamentally different in nature and require specialized technological approaches and architectures. Key considerations:

- Big data applications are “end-to-end”—typically involving pipelines of processing with steps that include aggregation, cleaning, and annotation of large volumes of data; filtering, integration, fusion, subsetting, and compaction of data; and, subsequent analysis, including visualization, data mining, predictive analytics and, eventually, decision making.

- Big data applications are characterized by greater diversity in software packages and software stacks, as well as a rapidly evolving software ecosystem—users should be able to run the software that best suits their application. Easy-to-use portals, or gateways, that hide the details of the underlying infrastructure from the end user are essential to the success of big data and extreme computing.

- Big Data systems will be heterogeneous in nature. The trends in industry, as well as the architecture of systems like Gordon and SDSC’s soon-to-be-deployed Comet, point towards a path forward where Big Data systems incorporate a range of capabilities matched up to the differing needs of different parts of a processing pipeline.
• Currently, Big Data systems in industry employ shared-nothing architecture with homogeneous, commodity components, to assist in manageability and scalability of the overall system. There is a separation between the active, stream-processing components of the system versus the analytics and post facto processing components. However, there is a desire to build “on-line everything” systems, where all processing could be done “inline,” with the data stream rather than as a post processing step.

From an interagency perspective, what do you think are the most high impact ideas at the frontiers of big data research and development?

There are a multitude of ideas at the frontiers of Big Data R&D with high societal impact. Investing in new technologies that support the ability to collect, analyze, and mine insights from extreme data sets has potentially high payoffs in climate science, cybersecurity, human genomics and personalized medicine, the “Internet of Things,” and financial market stability.

What new research, education, and/or infrastructure investments do you think will be game-changing for the big data innovation ecosystem?

In order to do meaningful innovation, researchers and developers need access to Big Data systems comparable in scale to problems being targeted. Google, Facebook and other commercial entities have deployed extreme scale Big Data systems, but these systems are not generally available for research & development beyond their respective companies’ proprietary needs. A game-changing investment would be one in extreme scale Big Data infrastructure (including the data sets) similar to the public investment in today’s supercomputing centers.

How can the federal government most effectively enable new partnerships, particularly those that cross sectors or domains?

New fields of research will emerge from the analysis of real world problems. A specific example is the NSF–funded WIFIRE project that analyzes real time events for wildfire prediction or the work underway at UC San Diego to apply temporal data models and techniques in the social sciences. In the latter, SDSC is at the forefront of computational sociology – a field made more powerful by utilizing parallel, semi-structured databases that operate in a cloud environment. Both project examples manipulate data to construct events and map the interrelationships between entities. What results is the emergence of innovation potential in the dimension of situational analytics. Another relevant example for connecting academia, industry and government is SmartCity San Diego (SCSD). Fundamentally SCSD is about understanding energy and modeling for predictive analysis, but also contributes to a variety of related industry partners and calls for innovation in data science: measuring, modeling integrating and interpreting models that inform projects across domains and business sectors. Funding cross-cutting projects analogous to SCSD would increase industry partnerships, lead to better informed city planning and have a direct impact on communities.

Rationale for inclusion in the strategic plan:

SDSC embodies a unique combination of computing at scale and deep data science expertise with a track record of innovation. Projects such as the NIH–funded NIF and IntegromeDB involve acquiring data, cleaning, transforming, creating ontologies and interlinking at a scale requiring SDSC’s computing infrastructure. SDSC’s innovation is in forming functional scientific applications and projects where Big Data tools and techniques can be effectively utilized. A specific example is NSF–funded OpenTopo, a massive store of LIDAR data in a parallel DB2 database with background
analytics that can span machines enabling scientific analysis on a large-scale platform. SDSC has the computational capacity to serve the confluence of domain scientists and technologies, innovating at their junction, especially where they don’t readily apply to each other – this is SDSC’s strength.

SDSC has a track record of educating undergraduate and graduate students as well as anticipating post-graduate training needs. A recent example is the NIH-funded BD2K, “An Open Resource for Collaborative Biomedical Big Data Training” that provides a learning environment with domain relevant data and software connected to pedagogically informed teaching materials on ‘best of breed’ cloud platforms. SDSC’s strength is in constructing such learning environments where many students can leverage a common, robust infrastructure.

**SHUG**

We are an elected committee of users of the national department of energy neutron source facilities, the Spallation Neutron Source and the High Flux Isotope reactor at Oak Ridge National Laboratory, who represent over 1200 users spanning domestic and foreign academic institutions, government laboratories, and US industry. We are both generators and consumers of Big Data, with individual experiments producing datasets too large to analyze on single machines or with existing software tools. We seek to extract scientific meaning and research and development value from the data generated by our experiments. The vast quantities of data produced by the advances in instrumentation and detector technology promise to usher in a new era in our ability to understand and design complex materials (which in turn are tied into ongoing initiatives for “Materials by Design” such as the White House Office of Science and Technology Policy Materials Genome Initiative). At the same time, we are not software or computer aficionados, and do not have the time or resources to devote to dealing with the challenges in the storage, transport, analysis and presentation of the large datasets produced in these experiments.

Our needs, as the scientists and researchers tackling problems from sustainable energy and curing cancer to fundamental science, would best be met if a national big data framework provided appropriate incentives to all parties so that the necessary hardware, software tools, and technical expertise are put in place to allow us to focus on the science and engineering of what we do, not on the manhandling of the data. This involves addressing many issues, including:

1) **Data Analysis Software**: The past forty years has seen the development of a wide range of bespoke software tools for the workup and analysis of data from national facility instrumentations – from codes that simply convert between different possible representations of the data to Rietveld codes for structural modeling. With rare exceptions, these codes were designed and built on the premise that data analysis would occur on a single computer, with manual intervention required to ensure the data input and outputs are in usable formats. While highly effective for their time, these tools are unable to cope with the volumes of data now available. Versions of these bespoke tools that use a cluster or distributed processing approach, with little or no manual intervention, simply do not exist. Yet without them, we as users are simply not able to tackle the science and engineering that we are interested in. A national big data framework must ensure that functional and usable (by non-computer expert) software tools exist.

2) **Data Storage and Transfer**: Today, even relatively modest datasets – 0.1 to 10 terabytes in size – present a significant challenge to move between the national facility and the home institution of the researcher, due to the patchwork nature of networking and storage infrastructure. Costly upgrades of national, facility, and home institution infrastructure are required to solve this issue, but typical federal grants do not provide separate funding resources to accomplish either networking upgrades or install and maintain the requisite data storage (especially at the home
institutions). To ensure a healthy data lifecycle, additional funds specifically directed to ensuring speedy data transfer and adequate data storage, in addition to those typically allocated for a grant, are required.

3) **Data Origin, Citation, and Preservation**: The data collected only has meaning when there is knowledge of the specimens/materials/etc. measured. Thus for data to be usable by those other than the researchers who originally carried out the measurement, appropriate annotation of measurement conditions (including as much sample information as practical) is critical. Further, it is simply no longer possible to include all of the primary data in a research publication or internal report, nor are there resources provided to ensure long-term availability of the primary (raw, unprocessed) data. To make the data part of the scientific record, and permit its usability by future researchers, mechanisms to cite raw (and processed) datasets, and preserve for “eternity” are needed. In part, this means moving the data to newer computing systems as they are created.

4) **Data Presentation**: For the integrity of the scientific and engineering process, it is critical that researchers be able to present, in a printed format, the characteristic features of the data that lead to the conclusions drawn from the data. This requires the development of new approaches to the presentation of Big Data analysis results. Achieving this also requires the development of new ways of visualizing and interacting with large data sets in their full glory, not just in reduced parameter space.

5) **User Training and Computing Capability Assets**: As with the computing revolution before it, there must be adequate training and investment in end-user usability of the software analysis tools for Big Data. Virtually all of the users have little of no programming experience – which locks them out of the tools currently available to handle large data sets in distributed fashion. Even for our users who are so inclined, often access to the necessary cluster computing facilities is either clunky or incurs a significant time delay, which impedes the research process.

In short, from our perspective as users, any national big data framework must focus on providing seamless access to data and analysis capabilities that do not require a computer science degree to utilize. At the same time, the tools and methods to tackle these issues cannot be developed independently of users - history is littered with examples of data processing software that failed because users were never brought in as active participants in the process and consequently either did not meet users' needs, or were so clunky and counterintuitive so as to effectively not exist. Instead any national big data framework must foster active engagement of users at all stages of tool development, while at the same time providing the necessary funding and stewardship for the maintenance of the tools so they can be used without distraction to tackle the grand challenges in health, science, and education.

Sincerely

Prof. Tyrel M. McQueen on behalf of the SHUG-EC

Chair, SNS/HFIR User Group Executive Committee (SHUG-EC)

Department of Chemistry

Department of Materials Science and Engineering

Department of Physics and Astronomy
Institute for Quantum Matter
The Johns Hopkins University
https://occamy.chemisyry.jhu.edu

SLAC_AMAB
Apurva Mehta and Amber Boehnlein; Staff Scientists, SLAC National Accelerator Laboratory

Ability to probe the chemistry and structure of materials at atomic scale has not only allowed a deeper understanding of fundamental processes in materials, but has illuminated knowledge-driven pathways to search the Materials Genome for novel new materials and devices, and the ability to scale-up and optimize them to address many of our technological challenges. With this vision in mind our nation has, over the last three decades, invested and continually upgraded x-ray, electron and neutron based national user facilities. With the continuing investments in brighter sources and development of faster detectors, the rate of data generation is perpetually accelerating; the pace of new discoveries and deepening scientific understanding, on the other hand, hasn’t seen equally dramatic increase. The widening gap arises because the data management infrastructure and advanced data analytics essential to convert data to knowledge hasn’t progressed sufficiently to keep pace with the volume as well as the changing modality of data collection. The widening gap between data and knowledge generation is also affecting the quality of the data: it is noisier, of uneven quality, and often does not optimally cover the most significant portion of the parameter space.

The traditional pathway to evidence based new knowledge is organizationally divided into sequential stages of a cyclic process. Traditional data to knowledge cycle begins with hypothesis generation, followed by experimental design, data collection, data quality assessment, and extraction of relationships and trends, which often leads to modification and refinement of hypothesis generation and experimental design. Historically, every stage of this cycle is curated and supervised by humans, often the same team of humans. Human eye and brain are superb at detecting subtle changes and seeing patterns buried in noise, but processing speed is a major limitation. As the speed of data generation increases, the human brain becomes overwhelmed and cannot keep up with on-the-fly curation, certification, and supervision, throttling the pathway from ‘Big Data’ to new knowledge. Accelerating pace of data generation, therefore, demands development of new tools and protocols, which overcome human limitations and even surpass their abilities, to reliably assess noise and spectral distortions and accurately extract subtle and hidden patterns from large, high dimensional and noisy data sets. Such tools must provide feedback within milliseconds (and with the commissioning of even higher throughput facilities must become even faster) to be useful for restore responsive control of data collection. In conclusion, the new data management software and hardware must be computationally efficient and fast, operate without human supervision and must be integrated back into the process of data generation to allow a knowledge cycle based on Big Data to operate optimally.

Our recommendation is then that National Big Data R&D Initiative must prioritize the development of scalable systems that seamlessly integrate the collection of experimental research data with sophisticated, smart and high-speed data analytics at high volume data (Big Data) generators, such as light sources and other national user facilities. This will insure that the highest quality data is collected under the most relevant experimental conditions and even the information hidden in noisy and distorted regions of data is reliably extracted, leading to accelerated pace of knowledge
generation. Federal investment that can integrate big data analytics with big data generation will enable high production of knowledge from the already significant investment of Federal funds in Big Data generating facilities.

Generation of knowledge at an accelerated pace from accelerating pace of data generation also requires close collaboration among researchers who collect experimental data; experts in statistics, image processing and computer science who assess the quality of data and extract hidden relationships and features from that data (data miners); and scientists who develop models based on extracted relationships, and predict yet undiscovered relationships, phenomena, and (device) behavior and functionalities. National investment that will enable integration of data collection and analytics (in a Big Data ecosystem) will effectively bring these currently distinct communities together. Training the next generation of scientists in these emerging cross-disciplinary techniques must explicitly be part of this agenda as well.

It is also critical that this new infrastructure is widely (and easily) accessible to researchers from broad range of scientific disciplines. This will enable diverse communities, from image processing to instrument developers, from electron microscopists to x-ray spectroscopists to learn and build from successes and failures of each other (e.g., x-ray scattering community building upon success of x-ray spectroscopy community, which in turn learns from successes in image processing and compression.)

National user facilities are ideal environment for the development of the proposed new big data R&D infrastructure as they are intellectual hubs and generate a significant fraction of big data in scientific research. Furthermore, their day-to-day operation is overseen by staff scientists who closely interact with the several thousand users from a wide range of disciplines and institutions (from academia to industry) and they themselves come from wide range of backgrounds and collectively have broad range of interests and knowledge.

To design an optimal knowledge generation engine based on Big Data requires three key new technical developments. The new developments include infrastructure to make data transparent and easily accessible by wide range of researchers, researchers beyond just those that generated the data. This necessitate development of common and self-describing data formats and data and meta-data archiving formats; and data storage and retrieval systems which are fast, secure, fail-safe, and easy to access and maintain.

It includes building machinery to assess the quality of the data faster than the rate of data collection; allowing sources of noise, artifacts and distortions to be discovered in real time and when possible to quickly alter the data collection conditions to optimize the data quality.

It includes building analytics tools – algorithms, protocols and approaches – that are robust, reliable, and computationally efficient and applicable to wide range of data as well as combination of different data streams (e.g., electron spectroscopy with optical spectroscopy; x-ray scattering with a chemical probe of reaction kinetics). Two types of data analytics are needed: 1) tools which are based on the knowledge of sources of noise and the detector responses and large population set (which enable use of statistical methods) to allow useful information to be extracted even from noisy and distorted data; 2) and tools that can, with minimal input from a human, reduce the Big Data to a small set of features and relationships that can not only be used to further curate the dataset, but can also be used to modify and refine data collection strategy on-the-fly (for example, collect more and closely spaced data across a phase transition).
Big Science in the Big Data Innovation Ecosystem

David Skinner LBNL and Amber Boehnlein SLAC

The work of designing, engineering, and operating the scientific instruments that allow us to observe and simulate the world in new ways is increasingly collaborative and data–driven. Big Science is one description for how large teams, devices, and data sets come together to produce scientific discovery and innovation. The trajectory toward team–based data–driven science is well documented by way of national strategy statements.1,2 Big Science is also well known through its big machines with big outcomes 3. Less–well communicated are the R&D agendas related to how the data from these instruments is acquired, analyzed, and shared. Here we connect the dots between these data–driven science instruments and the broader Big Data Innovation Ecosystem (BDIE).

How massive scientific research data figures into the broader ecosystem is a timely concern. Open Data memos and a general uptick in connected resources leads us to consider how the data flowing from Big Science instruments relates to emerging national–scale data frameworks. Below we identify areas for cross--sector partnership where Big Science has unique requirements and capabilities. The broader BDIE has much to offer Big Science as well.

Without defining what “Big Data” is, we can safely state that a deluge of scientific data presents challenges and opportunities that bear some overlap with broader data trends in the web and commercial contexts. Whereas data volume in content distribution scaling (Netflix, PB/year) and service scaling (Google, 450M users) are a well recognized aspect of the BDIE, challenges also arise where the velocity of a single data stream and the complexity of productive analytics are uniquely challenging. The “serial number zero” detectors and high--throughput techniques found in Big Science instrumentation require streaming data analytics and collaboration around large data sets at unique scales. R&D here informs next generation data science capabilities in the broader technology and workforce arenas.

We suggest there are three areas in which data–driven scientific instruments can, in the next five to ten years, broadly accelerate innovation. Facilities and frameworks for these three areas tackle tough problems that are currently unique to Big Science teams. Sooner than we think, these solutions will emerge as opportunities for technology transfer in manufacturing, engineering, education, and IT.

1) Streaming data analytics: Big Science detectors produce data streams at pace--setting bandwidths and modalities. We require both prompt analytics during operation as well as methodical re--analysis of data offline by large teams requiring advanced algorithms. Making decisions and discoveries from massive complex data streams is a Big Science specialty.

2) High--throughput workflow engines: Big science accelerates the process of scientific discovery and innovation through radical efficiencies in throughput. Found under the hood of the Human Genome Project, Super Nova Factory, and Materials Genome Initiative are high--throughput analytic engines that automate and organize tasks once assigned to individual grad students. Shrinking time--to--knowledge via high--throughput analysis combined with provenance and reproducibility demands present unique challenges.
3) Scalable software for teams: Big Science teams are large and complex, making decisions and discoveries based on large, diverse, and real-time data sets. The software and APIs in Big Science are unique in their scale. Software and policies that incentivize large teams to make effective use of Big Data toward shared curated models may serve as models for tomorrow's IT.

Data-driven scientific instrumentation arrives at these challenges head-on and often in advance of the broader ecosystem. Advancing the state of the art in the above areas or a superset of them is best done in partnership with the full spectrum of stakeholders. In analogy to the shared computing facilities, which emerged from widespread need for scientific computing in the 1970’s, today’s widespread challenges with scientific data are best served by reliable, efficient, and advanced facilities for scientific data.

Not all data challenges are alike. Across science teams, domains, and agencies the type and schedule of needs to be met will vary widely. An investment in a facilities approach to these challenges that identifies core data grand challenges provides translational leverage between these disparate requirements. The R&D that yields solutions to these challenges also educates tomorrow’s workforce at the cutting-edge of data science.

Big Science has a role in pioneering the frontiers of high-bandwidth streaming analysis, high-throughput workflows, and tools for distributed teams to collaboration. A scientific data facility, which addresses the above challenges, could directly attack grand challenges in research, leverage best-of-breed Big Data methods from the commercial world, and help train the next generation data scientist workforce. Interconnection of data from Big Science resources through a facility dedicated to that task is appealing to many stakeholders.

Let us not underestimate the aim and the need for such a super-facility in data. Big Science instruments bring new capabilities and understanding to the world. Discovery and innovation are often found where intercomparison of data from observations and simulations becomes scalable and efficient. A scientific data super-facility approach that promotes composable coordination of existing user facilities and sharing of scientific knowledge at Big Data scales is within reach. Inasmuch as we have recognized the transformative scientific value of intensely concentrated computational power through shared user-facilities, we must now pioneer strategies for intensely concentrated bandwidth and analytics.

1 http://science.energy.gov/about/


3 http://en.wikipedia.org/wiki/Nobel_Prize


Request for Input (RFI)-National Big Data R&D Initiative

Source: https://www.nitrd.gov/bigdata/RFI/02102014.aspx
A Notice by the National Coordinating Office, Networking and Information Technology Research and Development Big Data Senior Steering Group on October 2, 2014

Notice in Federal Register (https://www.federalregister.gov)

ACTION

Request for Input (RFI).

SUMMARY

This request encourages feedback from multiple big data stakeholders to inform the development of a framework, set of priorities, and ultimately a strategic plan for the National Big Data R&D Initiative. A number of areas of interest have been identified by agency representatives in the Networking and Information Technology Research and Development (NITRD) Big Data Senior Steering Group (BDSSG) as well as the many members of the big data R&D community that have participated in BDSSG events and workshops over the past several years. This RFI is a critical step in developing a cross-agency strategic plan that has broad community input and that can be referenced by all Federal agencies as they move forward over the next five to ten years with their own big data R&D programs, policies, partnerships, and activities.

FOR FURTHER INFORMATION CONTACT

Wendy Wigen at wigen@nitrd.gov or (703) 292-7921. Individuals who use a telecommunications device for the deaf (TDD) may call the Federal Information Relay Service (FIRS) at 1-800-877-8339 between 8 a.m. and 8 p.m., Eastern Time, Monday through Friday.

DATE

To be considered, written comments must be received by 11/14/2014.

SUPPLEMENTARY INFORMATION

Description

Overview

This RFI is issued under the National Big Data R&D Initiative. The NITRD BDSSG is inviting broad community input as it develops a National Big Data R&D Strategic Plan that can be referenced by participating agencies as they conceive and deploy their own big data programs, policies, and activities.
Background

The BDSSG was initially formed to identify big data R&D activities across the Federal Government, offer opportunities for agency coordination, and jointly develop strategies for a national initiative. The National Big Data R&D Initiative was launched in March 2012. Since the launch, the BDSSG has held numerous meetings and workshops, including a major showcase event of dozens of partnerships that will help advance the frontiers of big data R&D across the country. Many participating federal agencies have already established new big data programs, policies and activities and plan to do more in the future. Currently, the BDSSG is drafting a framework and establishing a set of priory goals for a National Big Data R&D Strategic Plan.

Objective

The major goal of this RFI is to gather information from multiple sectors to inform the development of an effective National Big Data R&D Strategic Plan. After the submission period ends and the feedback is analyzed, a workshop will be held to further discuss and develop the input received.

What We Are Looking For

As a big data stakeholder, we would like to 1) understand your particular role and point of view in the big data innovation ecosystem, and 2) using the draft The National Big Data R&D Initiative: Vision and Areas of Interest (My Note: See Below) as the basis, encourage you to provide comments and suggestions for how we might best develop an overarching, comprehensive framework to support national-scale big data science and engineering research and education, discoveries and innovation. Please include a description of the areas of critical investment (either within or across agencies), both currently and within a five to ten year horizon. Collectively, your comments could focus on one or more agencies, the set of NITRD agencies as a whole, or the national effort. Please keep in mind that the focus is on high level strategies for how agencies can leverage their collective investments. It will not focus on individual agency plans and will not contain an implementation plan. We are interested in all points of view on the activities that can best support research, development, and innovation in Big Data. However, we are not interested in specific research proposals or vendor offerings. While the NITRD agencies would welcome input on policies that are directly relevant to big data R&D, those policies that are more appropriately determined by the Administration or Congress (or both) are not relevant to this exercise.

Who Can Participate

This RFI is open to all. We especially encourage public and private sector organizations (e.g., universities, government laboratories, companies, non-profits) with big data interests to submit their ideas. Participants must also be willing to have their ideas posted for discussion on a public website and included with possible attribution in the plan.

Submission Format

All responses must be no more than two (2) pages long (12 pt. font, 1” margins) and include:

- Who you are—name, credentials, organization
- Your contact information
- Your experience working with big data and your role in the big data innovation ecosystem
• Comments and suggestions based on reading the initial framework, The National Big Data R&D Initiative: Vision and Priority Actions (My Note: See Below), and guided by the following questions:
  ◦ What are the gaps that are not addressed in the Visions and Priority Actions document?
  ◦ From an interagency perspective, what do you think are the most high impact ideas at the frontiers of big data research and development?
  ◦ What new research, education, and/or infrastructure investments do you think will be game-changing for the big data innovation ecosystem?
  ◦ How can the federal government most effectively enable new partnerships, particularly those that cross sectors or domains?

• A short explanation of why you feel your contribution/ideas should be included in the strategic plan
• Examples, where appropriate

In accordance with FAR 15.202(3), responses to this notice are not offers and cannot be accepted by the Government to form a binding contract. Responders are solely responsible for all expenses associated with responding to this RFI, including any subsequent requests for proposals.

The National Big Data R&D Initiative

Source: http://www.nitrd.gov/nitrdgroups/ma...ity_Themes.pdf (PDF)

PREDECISIONAL DRAFT September 26, 2014

Vision and Actions to be Taken

The following represents a preliminary, draft vision statement and possible actions to be taken as formulated by federal agencies participating in the Networking and Information Technology R&D Program Big Data Senior Steering Group (BDSSG) with input from external stakeholders. This framework outlines a vision and how agencies might move forward achieving this vision, both separately and collaboratively. Within these two thrusts, BDSSG agencies have identified some areas of high interest and potential impact.

Vision Statement

We envision a Big Data innovation ecosystem in which the ability to analyze, extract information from, and make decisions and discoveries based upon large, diverse, and real-time data sets enables new capabilities for federal agencies and the nation at large; accelerates the process of scientific discovery and innovation; leads to new fields of research and new areas of inquiry that would otherwise be impossible; educates the next generation of 21st century scientists and engineers; and promotes new economic growth. To this end, the NITRD agencies should consider how to most effectively:

• Create next generation capabilities by leveraging emerging Big Data foundations, technologies, processes, and policies
• In addition to supporting the R&D necessary to create knowledge from data, emphasize support of R&D to understand trustworthiness of data and resulting knowledge, and to make better decisions and breakthrough discoveries and take confident action based on them
• Build and expand access to the Big Data resources and cyberinfrastructure-- both domain specific and shared-- that are needed for agencies to best achieve their mission goals and for the country to innovate and benefit

• Improve the national landscape for Big Data education and training to fulfill increasing demand for both analytical talent and capacity for the broader workforce

The agencies will consider how to create new and enhance existing connections in the current national Big Data innovation ecosystem by, for example:

• Fostering the creation of new partnerships that cross sectors and domains
• Creating new gateways that enable the interconnection and interplay of Big Data ideas and capabilities across agency missions
• Ensuring the long term sustainability, access, and development of high value data sets and data resources

GovLoop

Source: http://www.govloop.com/

About Us

Source: http://www.govloop.com/about-us/

GovLoop’s mission is simple: connect government to improve government. We aim to inspire public sector professionals to better service by acting as the knowledge network for government.

GovLoop serves a community of more than 100,000 government leaders by helping them to foster collaboration, learn from each other, solve problems and advance in their government careers. GovLoop. We aim to inspire public sector professionals to better service by acting as the knowledge network for government. We do this through a variety of mechanisms including blogs, discussions, research guides, online trainings, mentorship programs, and more.

GovLoop works with top industry partners, such as ESRI, HP, Microsoft and IBM to provide resources and tools, such as guides, infographics, online training and educational events, for public sector professionals. GovLoop also promotes public service success stories in popular news sources like the Washington Post, Huffington Post, Government Technology, and other industry publications.

Resources

Source: http://www.govloop.com/resources/

About 130

Analytics

Source: http://www.govloop.com/resources/category/analytics/
WHY DATA IS YOUR MOST CRITICAL ASSET

Source: http://www.govloop.com/resources/data-critical-asset/

Patrick Fiorenza

September 30, 2014

Based on Mexico’s Yucatan Peninsula, the Mayan culture is home to one of Earth’s most unpredictable climates. The area is prone to both extreme drought and torrential rain. To prepare for prolonged periods without rain, the Mayans had to think of an innovative way to collect, manage and store their most precious asset: water.

The Mayans created an advanced system of reservoirs called chultuns and wells called cenotes. This system provided a consistent water supply to Mayans and allowed them to irrigate crops and access water year round.

Like the Mayans, government today must capitalize on its most precious resource: data. Leaders must understand how to harvest data and turn it into actionable solutions. In doing so, they can help transform our communities, improve our standard of living and create a more vibrant and engaged society. But government leaders must first understand the power of big data to change the business of government and provide deeper context and insights to enrich government services.

In this report, we will provide an overview of the current state of big data and the opportunity that big data analysis presents for government. Leveraging insights from the GovLoop community, we’ll seek to explore how agencies can capitalize on their most important asset: data. Specifically, our big data report explores:

- Government case studies from big data leaders
- Defines big data and insights on its applications
- Best practices to get started with data
- Interviews with public sector thought leaders on big data

GovLoop and IBM Guide Series Overview
In partnership with IBM®, GovLoop is conducting a four-part guide series, *The Journey to Big Data and Analytics Adoption*. Our report will emphasize how big data has transformed government agencies and help you set a road map for adopting big data solutions. Here’s what to expect in each chapter of our report.

- **Chapter 1: Leveraging Your Most Critical Asset: Data**
- **Chapter 2: Big Data’s Role in Countering Fraud**
- **Chapter 3: Improving the Safety of Our Communities**
- **Chapter 4: How Data Analysis Powers Smarter Care**

These four chapters will help in your journey to adopt big data analysis. Let’s start by exploring an important theme: why data is your most critical asset.

*The IBM Analytics Solution Center (ASC) is part of a network of global analytics centers that provides clients with the analytics expertise to help them solve their toughest business problems. Check out their [Analytics to Outcomes](http://www.govloop.com/resources/outcomes) group on GovLoop.*

**THE FOUNDATION FOR DATA INNOVATION: THE ENTERPRISE DATA HUB**


Patrick Fiorenza

September 18, 2014

Centrally managing data has long been a goal for IT managers. Now, with emerging technology and improvements to the way data can be stored and hosted, centrally managing data is finally a reality for organizations. As such, more and more agencies have been adopting an architecture – known as the enterprise data hub (EDH) – as the spot to consolidate and store data.

In this report, GovLoop explores the power of the EDH for public sector organizations. We spoke with the following four industry experts to help us understand how the EDH is transforming data management for government:

- Matthew Carroll, General Manager of 42six
- Joey Echeverria, Public Sector Chief Architect at Cloudera
- Erin Hawley, Director, National Security Programs, Cloudera
- Webster Mudge, Senior Director of Technology Solutions at Cloudera

With the [EDH](http://www.govloop.com/resources/rise-data-hub/), data can be stored in its original fidelity, integrated with existing infrastructures and supporting the flexibility needed for various kinds of workloads – such as batch processing, interactive SQL, enterprise search and advanced analytics.
Unlike in the past, today this can be done with the proper data protections, governance, security and management needs required by your agency. Powered by Apache Hadoop™, agencies can now capitalize on the ability to leverage their data in transformative ways.

Open publication - Free publishing

By adopting an EDH architecture, an organization has made the commitment to become information driven. This means they understand that data is the key to remaining economically viable in an increasingly competitive marketplace. "We even talk about [data] as a raw material, like steel and electricity," said Webster Mudge, Senior Director of Technology Solutions at Cloudera.

To learn about how your agency can capitalize on this raw material, be sure to read our latest report. My Note: See Below

Cloudera is revolutionizing enterprise data management with the first unified Platform for Big Data: The Enterprise Data Hub. Cloudera offers enterprises one place to store, process and analyze all their data, empowering them to extend the value of existing investments while enabling fundamental new ways to derive value from their data. Learn more here.

The Journey to Big Data & Analytics Adoption

Source: (PDF)

Introduction

"Within five years, the use of big data and analytics will just be a part of many agencies’ standard operating procedures, and eventually they won’t even think of it as big data. It will just be the way that they operate." Michael Stevens, Government market segment manager, IBM

In partnership with IBM®, GovLoop is conducting a four-part guide series, The Journey to Big Data and Analytics Adoption. Our report will emphasize how big data has transformed government agencies and help you set a road map for adopting big data solutions. Here’s what to expect in each chapter of our report.

IN THIS GUIDE

Chapter 1: Leveraging Your Most Critical Asset: Data

In this section, we will provide an overview of the current state of big data and the opportunity that big data analysis presents for government. Leveraging insights from the GovLoop community, we’ll seek to explore how agencies can capitalize on their most important asset: data.

COMING SOON
Chapter 2: Big Data’s Role in Countering Fraud

The second section of this report will explore how agencies can leverage data to improve fraud prevention methods. Growing use of the web for government transactions opens more areas to fraud. But with the proper tools and solutions, agencies can help combat fraud before it happens and be more efficient in service delivery.

Chapter 3: Improving the Safety of Our Communities

Agencies are looking at new ways to combat threats and keep communities safe. This means understanding how data can empower stronger insights and help prevent crime. With the use of analytics and robust data applications, agencies can improve the safety and economic vitality of their communities.

Chapter 4: How Data Analysis Powers Smarter Care

Governments are looking for ways to improve how they deliver services to citizens and take a more holistic view of an individual. This requires a thorough understanding of the role of data. This section will explore how agencies are taking innovative approaches to using data to improve service delivery and support communities of care.

These four chapters will help in your journey to adopt big data analysis. Let’s start by exploring an important theme: why data is your most critical asset.

Why Data is Your Most Critical Asset

“Right now something profound is happening: We are seeing a step change in terms of the kinds of volume, variety and velocity of data that is being produced worldwide.” – Tim Paydos, Worldwide director, public sector big data industry team, IBM.

Introduction

Based on Mexico’s Yucatan Peninsula, the Mayan culture was home to one of Earth’s most unpredictable climates. The area is prone to both extreme drought and torrential rain. To prepare for prolonged periods without rain, the Mayans had to think of an innovative way to collect, manage and store their most precious asset: water.

The Mayans created an advanced system of reservoirs called chultuns and wells called cenotes. This system provided a consistent water supply to Mayans and allowed them to irrigate crops and access water year round.

Like the Mayans, government today must capitalize on its most precious resource: data. Leaders must understand how to harvest data and turn it into actionable solutions. In doing so, they can help transform our communities, improve our standard of living and create a more vibrant and engaged society.

But government leaders must first understand the power of big data to change the business of government and provide deeper context and insights to enrich government services.

Across the web, you can find thousands of ways to define big data. It’s often defined as datasets that are too large and complex to manipulate or interrogate with standard methods or tools.
Tim Paydos, Worldwide director, public sector big data industry team at IBM, took a different angle on defining big data – by focusing on what it’s not.

“Big data is not a technology; it’s an event, it’s an occurrence,” Paydos said. “It’s a phenomenon, it’s something that is happening around us day to day, and it creates a challenge and an opportunity. So how do you rise to the challenge and seize the opportunity?”

One common definition of big data involves using the four V’s:

- **Volume**: The quantity of data that agencies now collect.
- **Velocity**: The speed at which data is created.
- **Variety**: The various data types that agencies have access to.
- **Veracity**: The authoritative nature of government data.

“The way I like to think about the value of big data is that it’s enables government to move from a stance of responding – to social issues, to crime, to threats, to fraud -- to one of anticipating these things,” said Michael Stevens, government market segment manager of IBM’s Big Data & Analytics Category team. “It gives governments the ability to anticipate things and thereby deal with them in a much more efficient manner. It has the capability to change the way the government does business, from an efficiency standpoint, from an effectiveness standpoint, from an outcomes standpoint.”

“People think of big data as this big scary thing for which they’re going to have to completely rip out their existing infrastructure and put in new technology,” Stevens added. “Nothing could be further from the truth. This is particularly true for local governments who don’t have the resources to perhaps do some of the things that defense or national security agencies are doing. Agencies can start with a small project and build out from there.”

For agencies to capitalize on big data, officials must understand that just as mission requirements and business needs change so will the opportunities with big data.

"I truly believe that big data is going to be transformational and revolutionary, but the great irony is that the path to harnessing big data is iterative and evolutionary,” Paydos said.

From our interviews with thought leaders, we’ve found that big data is:

- **Enabling** government to capitalize on the volume, velocity, variety and veracity of government data.
- **Creating** a more proactive government.
- **Improving** trust and government services.
- **Creating** a more efficient government, capable of meeting complex demands.

Increasingly, data is the key to transforming our government and now is the time to create the proper infrastructure.

**How IBM Can Help**

IBM’s services transforms agencies by helping them capitalize on data.

“One of the things we do and specialize in is working with organizations on how they get started [with big data],” Stevens said. “I recommend starting with bite-size pieces, existing technology and existing data.”
By starting small and not reinventing the wheel, IBM helps organizations create their big data blueprint and work through the common cultural and information technology challenges to adopting big data.

“One of the clients I worked with recently said to me that this was the first time he ever looked forward to the weekly update meetings, because he knew he’d always see progress and always learned something as a part of that, so that was one of the best compliments I think a client services persons can get,” said Brian Murrow, associate partner in the strategy & analytics unit at IBM and leader of IBM’s federal financial regulatory consulting practice.

IBM®’s case studies extend across all verticals and they include all sectors. We’ve highlighted six compelling case studies of big data use and their outcomes.

Analytics 101: Defining Different Kinds of Analytics

Several kinds of analytics are shaping the public sector and powering big data initiatives. Before you adopt big data, you must understand them so that when you start a big data program, you know what kind of analytics approach to take.

Some examples include:

• Business intelligence (BI) analytics, which helps decision-makers easily receive access to information and improve reporting and analysis. BI works to provide information at the right time to the right people. For instance, an emergency response team can benefit from BI analytics by receiving real-time data on individuals’ location, weather or facilities.

• Performance management analytics, which helps managers define strategy, understand the workforce and improve reporting and transparency. Organizations can track metrics, key performance indicators and other data to drive mission success.

• Predictive analytics, which includes a variety of statistical techniques to model data and analyze historical data, such as social program or crime data to anticipate social needs or predict crime.

• Entity analytics, which focuses on improving the accuracy and consistency of data across disparate datasets by resolving conflicts in records. For instance, a criminal might go by the name Joe, Joey or Joseph across datasets. Entity analytics will connect the dots between the data to form a common profile of a criminal.

• Descriptive analytics, which typically deals with structured data and mines historical data to categorize, classify and group information.

• Content analytics, which uses enterprise content management services to unlock data that is trapped in documents. This enables organizations to identify new insights and quickly disseminate information, for example, improving the way they investigate and prevent crime.

• Social media analytics, which provides a constant stream of information. For example, a police department can leverage the highly unstructured data to find photos, locations or a track down witnesses from crimes, all by monitoring social media services.

• Intelligent video analytics, which can help identify events and spot patterns of behavior through video analysis of locations. Video analytic software can monitor events in real time, send alerts and automatically spot specific incidents and patterns.

Big Data Case Studies

Across all sectors, big data is changing the way society operates. The private sector has found many ways to predict demand and identify insights by leveraging big data. And many public-sector institutions are doing just the same. We
highlight six case studies of public and private organizations that are using big data in compelling ways. University of Texas.

**Big Data for Medical Breakthroughs**

**How the Anderson Cancer Center applies big data to cancer research**

At the University of Texas MD Anderson Cancer Center, oncologists have discovered a new method to help improve the treatment of cancer patients. By analyzing patients' DNA, oncologists can gather insights into how to best provide treatment. The research literature on cancer treatment is rapidly growing, and oncologists needed a way to map patients’ history to emerging medical literature to tailor their treatment.

Using IBM’s Watson, an artificially intelligent computer system, oncologists can now analyze and understand published literature and make improved recommendations on care based on the latest research, helping to improve survival rates and provide smarter care to cancer patients.

**Predicting Demand with Analytics**

**Lessons learned from Blizzard Ski’s ability to predict ski conditions with data**

The owners of Blizzard Ski, an online store for skiing equipment, must order the right amount of supplies to build nearly 400,000 skis per year – skis that often require 18 different materials. To help with the planning and purchasing of those materials, the company has started to use data.

Blizzard Ski has created forecasting models built around weather patterns, ski trends and any short-term shifts in their market that may affect demand for skis. As the company has become more data driven, it has been able to improve inventory and become more agile in the development of skis, reducing production time by nearly half.

**Improving the Environment**

**How the city of Dubuque, Iowa, reduced its carbon footprint**

Officials in Dubuque, Iowa, were looking at ways to reduce the city’s carbon footprint and decrease their water consumption. Integrating cloud technologies and analytics from IBM, they deployed a solution for real-time sustainability monitoring, enabling a full view of energy management.

“Monitoring water consumption every 15 minutes, the system securely transmitted that anonymous data along with information on weather, demographics and household characteristics to the cloud, where it was analyzed,” an IBM report states. “The system also quickly and automatically notified households of potential leaks and anomalies — and provided citizens with a better understanding of consumption patterns.”

The system’s pilot test had great returns for the city: Water use decreased by 89,090 gallons among 151 households, which is equal to a 6.6 percent reduction of water usage. The pilot also predicted that community-wide, Dubuque could decrease the water use of 23,000 households by 64,944,218 gallons and increase leak detection by 8 percent citywide.
Dubuque is a great case study of a local government using analytics to improve the quality of services provided to citizens.

**Keeping Communities Safe**

**How Miami-Dade County combated crime**

The Miami-Dade Police Department is one of many agencies leveraging analytics and data to build safer communities in the Digital Age. Miami-Dade is Florida’s largest county, with 2.5 million citizens, and it’s also the seventh largest county in the United States. As is the case with many counties in Florida, tourism is one of its largest economic drivers. To promote it, officials have placed an emphasis on public safety.

To that end, the department turned to Blue PALMS (Predictive Analytics Lead Modeling Software), an advanced analytics model helping to provide stronger leads for the county police department’s Robbery Investigations Section. The model allows the agency to quickly and efficiently allocate resources to better combat robbery cases and reduce street crime. Additionally, Blue PALMS lets officials analyze cold cases against historical data to help identify leads and discover previously unknown insights to help in investigations.

**Improved Health Services to Communities**

**How Alameda County has transformed health care with data**

Alameda County is a community of 1.6 million in California. The Alameda County Social Services Agency (SSA) employs a staff of 2,200 who provide assistance to 11 percent, or 176,000, of Alameda residents. SSA delivered services ranging from employment assistance to help for the homeless to elder care. Before adopting IBM’s Smarter Care approach, which helps organizations take a holistic view of citizens, each service delivery system was run through a separate office and employees had to manually enter and track data across systems, leading to errors in data and outcomes. In addition, 1,200 caseworkers were responsible for 500 to 600 cases at any given time.

To overcome these challenges, SSA officials decided to create a single platform for tracking customer data. Specifically, they wanted a software solution that would increase employee productivity and make their system transparent and accountable.

The Alameda County SSA deployed two programs simultaneously: the Social Services Integrated Reporting System and IBM InfoSphere Identity Insight with IBM Cognos and IBM InfoSphere Warehouse. These systems delivered capabilities in advanced entity analytics, business performance and integrated data warehousing. Ultimately, they transformed SSA services from inefficient and siloed to consolidated and holistic. The transformation took only six months. As a result, the agency gained a return on investment of 631 percent in a two-month payback period and an average annual benefit of $24.7 million.

**Fighting Tax Corruption with Data**
How New York leverages big data to reduce tax fraud

In New York state, Nonie Manion, executive deputy commissioner of the Department of Taxation and Finance, leads a team of 1,600 auditors, all unified by a common mission: reduce tax fraud to preserve fairness for citizens. Manion is leading the charge to reduce waste, fraud and abuse by better understanding data and tracking issues.

Manion championed a predictive analytics initiative to detect questionable refunds before they are dispersed to citizens. In doing so, the office is able to protect critical tax revenue and use that funding more effectively. Since it started using predictive analytics, the office has decreased the revenue drain caused by questionable refunds by $1.2 billion.

Realizing the Promise of Big Data

We live in a society bursting at the seams with data, and public organizations must not let this resource go to waste. Big data analytics provides the opportunity for governments to leverage data and make evidence-based decisions.

In a recent report, Realizing the Promise of Big Data: Implementing Big Data Projects, the IBM Center for the Business of Government and Arizona State University Professor Kevin Desouza explored the emerging world of big data and its potential for revolutionizing the public sector. My Note: Have Wiki Version of This

Big data is a buzzword that has increasing popularity, but is not always understood, they said.

"Big data is an evolving concept that refers to the growth of data and how it is used to optimize business processes, create customer value and mitigate risks," Desouza said. "For public managers, big data represents an opportunity to infuse information and technology into the design and management of organizations, personnel and resources."

Desouza believes the most challenging big data “V” for the public sector is variety. Many organizations struggle to find the right methods of integrating data from newer platforms into legacy systems.

Although extracting meaningful information from legacy systems is a challenge, there are many positive examples of big data adoption in the public sector. In the report, Desouza explores cases from the U.S. Postal Service, Internal Revenue Service, the state of Massachusetts and more.

He delves into a particular case study involving New York City and how local municipalities use datasets to enhance management of transportation, utility and infrastructure systems. New York City’s Business Integrity Commission (BIC) administers licenses to sanitation haulers and wholesalers. The agency regulates more than 2,000 businesses citywide and must ensure that waste is collected, hauled and disposed of properly.

The increased value of recyclables and yellow grease (cooking oil) – which can be made into biodiesel fuel – caused a surge in criminal waste collection; unlicensed individuals were illegally collecting and selling these products for a profit. To stop this behavior, BIC and the Mayor’s Office of Data Analytics collaborated with the departments of Health and Mental Hygiene and Environmental Protection to leverage industry data on grease production, restaurant permits and sewer backups. By cross-referencing this data, the agencies created a yellow grease heat map to identify potential hotspots of unlicensed waste collection activity. Since implementing this technology, BIC reported a 30 percent increase in discovered violations and a 60 percent decrease in manpower for grease enforcement.
As this example shows, big data analytics can revolutionize public-sector management. Still, in a survey of chief information officers at all levels of government, Desouza identified 10 findings that speak to the current state of big data in the public sector. Some of the major ones indicate “public agencies are in the early days of their big data efforts” and not many CIOs “anticipate significant investments in technology.” On the other hand, CIOs believe that collaboration, leadership and working groups are necessary for additional big data programs to be adopted. Desouza notes that CIOs are “becoming champions of analytics and evidence-driven decision-making.”

The report concludes with a set of key steps and best practices to overcome the unique challenges public officials face when implementing big data projects. The implementation process is broken into three stages. The first is planning. CIOs must do their homework, seek out peers’ expertise and ensure that the project is aligned with other agency efforts. Furthermore, it is important to develop process and outcome measures and assess risk and privacy concerns before starting a data project. In other words, begin with the lowest-hanging fruit. Desouza observes that small opportunities are the easiest to tackle before moving on to larger tasks.

The next stage is execution, the most crucial aspect of which is communication. CIOs should constantly check in with all parties involved and gauge the pulse of the program. It is important to remain focused and manage scope creep.

The final stage is post-implementation. This involves conducting postmortem analysis and determining the impact of the project. CIOs can assess the lessons learned and identify the next project.

Big data is becoming more pertinent for public managers. Agencies will continue to wrestle with integrating big data analytics into their current IT systems and strive to make data-driven decisions. The IBM Center for the Business of Government’s “Realizing the Promise of Big Data” report offers helpful resources for public organizations navigating the path to implementing successful big data projects.

10 Questions to Start Your Big Data Initiative

Below is a series of questions you can use to spur a big data conversation at your agency.

1. What problem are we trying to solve? How can data help?
2. What kind of data do we need access to?
3. Who are the main stakeholders and how do we engage them?
4. How are we going to track, assess and monitor progress?
5. Can we pilot test a few programs and start small? What can we learn from starting small and building out?
6. What does our workforce look like? Do we have the right workforce in place?
7. Is our entire organization/department aware of the strategies? How have we engaged people in the process?
8. What are our feedback mechanisms?
9. Are we delivering on the needs of our users? How do we know?
10. What best practices have we learned from our peers and similar projects? What roadblocks might occur, and how do we stop them?

These 10 questions are just the start for your agency. They provide a great framework to help you to start thinking about how to use data in new and innovative ways.

9 Steps to Starting Your Big Data Journey

GovLoop and IBM have identified nine key pointers that will help you get started on your path to using big data.

1. Master the Art of the Possible

Organizations’ leaders must be able to cast a vision of how to use analytics. This means they must have a firm grasp on existing data, agency mission and what improved use of data can accomplish. Without mastering the art of the possible, agencies will struggle to identify proper data applications and use cases.

2. Don’t Reinvent the Wheel

Thousands of agencies are conducting analytics programs. Reach out to your peers. They can provide you guidance and insights to help drive change. Also, be sure to share your success and lessons learned. This is valuable intelligence that can help your peers in government.

Additionally, our subject-matter experts advise organizations to consider the IT infrastructure and data that they already have within the agency, so they are not starting from scratch.

“Start with existing data, so you don’t go down a rabbit hole of looking for perfect data,” Murrow said. “By starting with existing data, you can achieve near-term results.”

3. Show the Business Value

Everything you do revolves around business value. If your organization has core metrics or deliverables you must meet, show how analytics can help and exceed those goals.

“One best practice is to have a point of view on an enterprisewide analytics blueprint and figure out what this could look like for your agency,” Murrow said. “It doesn’t have to be a complicated blueprint, just jot down a series of ideas around pilots and what you might want to do using big data and analytics. It’s important to think about what will help advance your agency in terms of analytics and big data maturity.”

4. Select a Mission-Critical Problem to Solve

To start your big data journey, you should first articulate a clear mission need or result. The case studies show that when you select a clear problem and have access to authoritative data, you are well on your way to making data work for your agency.
“When developing your blueprint, think of the current business priorities of your organization’s leadership, and you need senior-level support as well, so start with those current business opportunities and figure out what the measurable business value is for those, so you know you are achieving your objectives,” Murrow said.

5. Assess Capabilities against Requirements

Once you select a program, it’s essential to gain an understanding of the data and infrastructure that is already available within the agency. This will allow you to leverage current capabilities and cut costs by not purchasing expensive infrastructure.

It’s also important to remember that not all data sets are created equal. When thinking about your data, analyze high-priority data vs. low-priority data. For example, you may be able to save on cloud storage by optimizing data sets in the cloud and leveraging on-premise storage. If you know which data is essential to power your agency, you can more effectively manage your organization’s information.

“Take an inventory of what your data assets are both within your agency and enterprise and across government, outside your agency and even within the commercial realm of data that is available,” Stevens said. “And then assess your current architecture and capabilities against what’s required to achieve those goals.”

6. Start Small and Have Clear Outcomes

Analytics is a very broad subject, and there are multiple ways to use data to solve organizational issues. Start small and think about a few core mission projects. Use these pilot programs as an incubator, then scale out to larger initiatives.

“Many agencies are going to be starting this from scratch; they’ve never done it before. So, start small, choose a use case that is likely to succeed, execute on it, review it, make adjustments and then move on to other use cases,” Stevens said.

7. Collaborate Across Government

The data that you have in your agency can be valuable to other departments. It is imperative on your data journey to connect with your peers and have a firm understanding about the kinds of data you have within your agency, and what data you may need to advance your agency.

“Explore which assets could be made open and available to the public to help spur innovation outside the agency,” Stevens said.

8. Measure Impact and Develop ROI

One of the keys to big data is creating the proper mechanisms to assess and measure impact. Developing the right metrics is essential to big data success.

“Another area that is really important is looking at the total cost of ownership,” Murrow said. “To help you determine which pilot to do, you really want to look at total return on investment and total cost of ownership. This isn’t a science fair project. If you can’t tie it to the mission, then maybe you should really be rethinking how you are using big data and analytics.”
9. Make Data Actionable: Tell Case Studies

To gain support across your agency, develop case studies and great use cases. When you can tell a story of how money was saved or more services were delivered or how data-based decisions led to meeting goals, you will be able to gather more support from your peers.

Data offers so much opportunity to transform the business of government. By following these steps, you will be able to leverage data in new ways to reimagine how your agency delivers services.

Resources and More Reading

About GovLoop

"There is a very big impact, as all organizations look at some of the agencies that have been successful and learn how to extrapolate that to their organization." Brian Murrow Associate partner in the strategy & analytics unit, IBM

GovLoop’s mission is to “connect government to improve government.” We aim to inspire publicsector professionals by serving as the knowledge network for government. GovLoop connects more than 140,000 members, fostering cross-government collaboration, solving common problems and advancing government careers. GovLoop is headquartered in Washington, D.C., with a team of dedicated professionals who share a commitment to connect and improve government.

For more information about this report, please reach out to Patrick Fiorenza, senior research analyst, GovLoop, at pat@govloop.com.

GovLoop

1101 15th St NW, Suite 900
Washington, DC 20005

Phone: (202) 407-7421 Fax: (202) 407-7501

http://www.govloop.com

Twitter: @GovLoop

Recognition

Thank you to IBM for their support of this valuable resource for public-sector professionals.

Author: Patrick Fiorenza, GovLoop’s senior research analyst

Designers: Jeff Ribeira, GovLoop’s senior interactive designer, and Tommy Bowen, GovLoop’s junior designer.

Editor: Catherine Andrews, GovLoop’s director of content
About IBM

The world isn’t just getting smaller and flatter, it is also becoming more instrumented, inter-connected and intelligent. As we move toward a globally integrated economy, all types of governments are also getting smarter.

IBM® provides a broad range of citizen centered solutions to help governments at all levels become more responsive to constituents, improve operational efficiencies, transform processes, manage costs and collaborate with internal and external partners in a safe and secure environment.

Governments can leverage the unparalleled resources of IBM through IBM Research, the Center for the Business of Government, the Institute for Electronic Government and a far-reaching ecosystem of strategic relationships.

IBM®, the IBM logo, ibm.com, IBM® InfoSphere® Identity Insight, IBM® Cognos®, IBM® InfoSphere® Warehouse are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. If these and other IBM® trademarked terms are marked on their first occurrence in this information with a trademark symbol (® or ™), these symbols indicate U.S. registered or common law trademarks owned by IBM® at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM® trademarks is available on the Web at “Copyright and trademark information” at: http://IBM.com/legal/copytrade.shtml. Other product, company or service names may be trademarks or service marks of others.

Resources and More Reading

Realizing the Promise of Big Data [IBM Report]

http://www.govloop.com/community/blo...ta-ibm-report/

Analytics in Action: City of Dubuque Case Study

http://www.govloop.com/analytics-in-...ue-case-study/

To Face New Threats, We Need New Ways of Thinking

http://www.govloop.com/to-face-new-t...s-of-thinking/

Alameda County – Business Values to Social Services

http://www.govloop.com/alameda-count...cial-services/

Using Data Analytics to Counter Fraud: New York State Tax Case Study

Incorrect URL

MD Anderson and IBM Watson Collaborate to End Cancer

https://www.youtube.com/watch?v=P7Sh...708A8&index=52
Introduction

Centrally managing data has long been a goal for IT managers. Now, with emerging technology and improvements to the way data can be stored and hosted, centrally managing data is finally a reality for organizations. As such, more and more agencies have been adopting an architecture – known as the enterprise data hub (EDH) – as the spot to consolidate and store data.

In this report, GovLoop explores the power of the EDH for public sector organizations. We spoke with the following four industry experts to help us understand how the EDH is transforming data management for government:

- Matthew Carroll, General Manager of 42six
- Joey Echeverria, Public Sector Chief Architect at Cloudera
- Erin Hawley, Director, National Security Programs, Cloudera
- Webster Mudge, Senior Director of Technology Solutions at Cloudera

With the EDH, data can be stored in its original fidelity, integrated with existing infrastructures and supporting the flexibility needed for various kinds of workloads – such as batch processing, interactive SQL, enterprise search and advanced analytics.

Today this can be done with the proper data protections, governance, security and management needs required by your agency. Powered by Apache Hadoop™ (see Figure 1), agencies can now capitalize on the ability to leverage their data in transformative ways.

By adopting an EDH architecture, an organization has made the commitment to become information driven. This means they understand that data is the key to remaining economically viable in an increasingly competitive world.

“We even talk about [data] as a raw material, like steel and electricity,” said Webster Mudge, Senior Director of Technology Solutions at Cloudera.

But to leverage this raw material to improve decision-making, organizations must overcome numerous obstacles. One of the major challenges is centrally managing data from disparate locations and in various file types.
That’s why so many agencies are looking to the EDH as a means to centralize their data. In doing so, agencies have witnessed significant reductions in the costs of data management – partly due to the reduced duplication of data – and can now have information available to multiple users. The key to the EDH is that it is an open, scalable data architecture that is shared by multiple computing frameworks. This architecture is what has separated EDH from previous data management models. The EDH is driving many benefits for government agencies, such as:

1. Shared resources, systems, memory and data sets.
2. Minimal data movement.
4. Very straightforward data acquisition
5. Centralized multipurpose workloads.
6. Centralized processing.
7. No incremental costs for new computing.
8. Easy provision of new data.

These benefits are helping agencies to unlock new insights from information. But an often-overlooked benefit of the EDH is that the architecture consolidates not only contemporary data, but also historical data. This gives government the ability to capture data in real time and measure against former trends. This is because the EDH architecture allows organizations to quickly and easily integrate with existing legacy systems.

“What you’re able to do with the enterprise data hub is bring in data sources over time,” said Joey Echeverria, Public Sector Chief Architect at Cloudera.

“[With an EDH,] legacy applications that connect directly with relational databases can also connect directly to an EDH, so that you can transfer data in the places where they exist now into an EDH, and integrate with the existing tools that you use for accessing that data.”

With the ability to collect multiple kinds of data from disparate systems, the EDH is a powerful tool to help agencies unlock insights from their data.

With the EDH, now the computing power is being brought to the data, improving application development, access to software and information sharing.

“Before this EDH architecture you had different computing environments and you were moving the data around from one application to the next, moving data to the compute,” said Mudge. “The fundamental shift here is that the data is now the center, and the computing is being brought to the data.”

“There are obvious cost advantages in centralizing your management of data, both from de-duplication, and just from being able to use cost effective platforms, without scaling at such steep cost curves,” said Echeverria.
The solution at the center of an EDH is Apache Hadoop, a 100% open source platform to store and process data. Hadoop removes the need for organizations to rely solely on expensive, proprietary hardware to store and process data. Yet Hadoop alone lacks the proper governance, data protection, and management solutions needed for the public sector. With Cloudera’s commitment and investments in the open source community, many challenges - such as previously mentioned governance, auditing, data protection, and centralized management with Hadoop - have been addressed.

Two examples of the solutions that Cloudera has created to overcome solution gaps in Hadoop are Cloudera Navigator and Cloudera Manager. Both of these have been specifically designed to complement Hadoop and the EDH as a whole. These solutions are providing government agencies with a full suite of solutions needed to capitalize fully on the power of Hadoop. With Cloudera Navigator and Manager, Cloudera has become the lead engineering force behind Hadoop-powered analysis and storage.

Cloudera’s enterprise data hub offers the ability to:

- Meet compliance regulations by providing immediate access to data and archiving content digitally.
- Complement existing data warehouses to help manage costs and improve performance.
- Support a movement towards self-service business intelligence within the agency.
- Offer a consolidated approach to searching across data systems.
- Give you advanced analytic solutions that can quickly and efficiently predict and spot fraud and anomalies.

Facilitating Converged Analytics & Data Governance

The EDH is also facilitating an effective and efficient data ecosystem and is essential to supporting the adoption of converged analytics strategies. The basis of converged analytics is to provide an organization the ability to gain a holistic view of its data by employing many different kinds of computing simultaneously against a single shared data set without duplication or data movement.

“[Converged analytics] is really critical, because you just don’t know and can’t dictate what exactly is going be the right mix of analytics at that time,” said Mudge. “And you don’t necessarily want to dictate, because that’s going to get in the way of actually solving the problem.”

To truly leverage data today, organizations must be able to aggregate different data types, run reports against data, all while maintaining data integrity and fidelity.

Yet, with this kind of data freedom, administrators are challenged on how to best maintain the right level of business control. To overcome these trials, Cloudera has created Cloudera Navigator, a native data management application for Hadoop. Navigator is designed to provide data management capabilities for users, administrators, and auditors, helping them to govern and explore the data within the EDH. Specifically, Navigator is guided by four core principles, which Cloudera defines as:

- DATA AUDIT AND ACCESS CONTROL – Verify appropriate user/group permissions and report on data access for files, records and metadata.
• DISCOVERY AND EXPLORATION – Find out what data is available and how it’s structured so that you can use it effectively.
• LINEAGE – Trace the progression of data sets back to their original sources to verify reliability of results.
• LIFECYCLE MANAGEMENT – Ensure the correct placement and retention of data based on value or policies.

Cloudera has also created Cloudera Manager, a system management application for Hadoop and the enterprise data hub. Cloudera Manager provides an agency the ability to deploy, configure, and manage an enterprise data hub, ensure quality of service and performance, and streamline operations – all at scale. Cloudera Manager:
• Provides a cluster-wide, real-time view of nodes and services running.
• Provides a single, central console to enact configuration changes across your cluster.
• Provides per-node services templates for rapid provisioning and rolling updates to maintain service level agreements (SLA) and quality-of-service (QoS) measures.
• Incorporates a full range of reporting and diagnostic tools to help you optimize performance and utilization

With the capabilities of Cloudera Manager, IT operators get end-to-end administration for Hadoop and their EDH. Ease-of-use is critical for success with distributed systems like Hadoop, especially as the platform grows to handle new data and new usage.

“There are obvious cost advantages in centralizing your management of data, both from de-duplication, and just from being able to use cost effective platforms, without scaling at such steep cost curves.”

Joey Echeverria, Public Sector Chief Architect at Cloudera

EDH In Action: Defense Intelligence Agency Case Study

The Defense Intelligence Agency (DIA) provides military intelligence to warfighters, defense policymakers and force planners to the Department of Defense and the Intelligence Community. The DIA is responsible for the planning, management and execution of intelligence operations during peacetime, crisis, and war.

Tasked with this mission, it is essential that the DIA has the ability to provide people information safely, securely and efficiently, no matter what the situation. This comes as quite the challenge, as the agency:
• Manages 400 plus apps within the enterprise.
• Has over 1,000 active data sources consuming data on the order of terabytes daily.
• Supports over 230,000 daily users with mission and business needs.
• Has a network that’s deployed worldwide at every combatant command.

At the DIA, Matthew Carroll, general manager of 42six, which is a CSC company, led the adoption of an EDH. During a recent online training with Gov-Loop, Carroll provided the following statistics on the early return on investments the DIA has witnessed:

Distributed query engine is able to cut duplication in storage, cut down data explosion by 100x fold, expecting a savings of $15-20M per year.
• Deployment model and SDK has cut down time to delivery of prototypes on the cloud from six months to thirty days.
• Applications can be accredited within a week from an average of four to six months.

https://semanticommunity.info/Data_Science/Data_Science_for_the_National_Big_Data_R_and_D_Initiative
Updated: Sun, 14 Jul 2019 10:27:14 GMT
Powered by mindtouch
• Systems and security personnel can be removed from contracts forecasting a savings between $25-35 million per year.
• Re-use of existing database investments has decreased transition costs between $20-30 million over the next two years.

Although these are significant improvements, the DIA faced numerous challenges to get to this point. Carroll noted that the agency had to navigate the budget to transition over 400 custom apps. App migration is not an easy task, and they also needed more time spent on security efforts.

The EDH provided the necessary framework and architectures to help the DIA transform how they achieve their mission.

“The EDH is creating the necessary interfaces for specific tools for each use case,” said Carroll. “And it contains the ability to aggregate all the data in a way so that it becomes discoverable, and indexable in a way that makes sense for application developers – without fundamentally changing the data itself, promoting reuse and rapid development for multiple applications.”

**Figure 2 Monitoring at the Tactical Operations Center**

Source: [https://www.flickr.com/photos/soldierter/3992505713](https://www.flickr.com/photos/soldierter/3992505713)

**Figure 3 Searching Documents**

Source: [https://www.flickr.com/photos/soldierter/13995623764/](https://www.flickr.com/photos/soldierter/13995623764/)

**How Can You Get Started**

In nearly every government program, data has been a key factor to deliver better outcomes.

“Agencies have leveraged [the EDH] across the board, whether it’s the intelligence community, Department of Defense, or civilian agencies,” said Erin Hawley, Director, National Security Programs, Cloudera.

In particular, Hawley identified an example from one client from the defense community. Tasked with need to archive flight mission data, the agency had invested significant hours physically going to archives facilities to find previous mission data records, and bring the information back to the team for manual entry.

“It was so expensive for them to do it that way, it was cheaper to fly the missions again,” said Hawley.

By leveraging an EDH architecture, the agency was able to store its archival data at minimal cost, yet also make the data immediately available for analysis and avoid the hidden costs of data access that they faced before. This case study is a testament to the radical shift that agencies are going through in terms of data management.

To start thinking about how your agency your can leverage an enterprise data hub; here are three key starting points.
Ask the Right Questions

One of the keys to success with the EDH is that agencies must work hard to understand how data can improve their business operations. This requires leadership and a commitment to becoming a data driven agency. Preliminary questions that organizations should explore include:

- What's important to our business operations? Our mission operations? Where is there overlap and separation?
- What kinds of data are we capturing? What kinds will we have to capture?
- Can we identify data sets that might have value to the organization if viewed, analyzed, or correlated differently?
- What kinds of insights from data will help us achieve our mission?
- What's the current status of data in our agency? Is our culture data- or information-driven?

By starting with these questions, agencies can develop deeper insights around data. "We’ve seen organizations combine datasets that previously we thought to be completely unrelated, and they end up being able to find new insights," said Echeverria.

Start Small and Iterate

Successful organizations have tackled a specific problem through data management, and then reached deeper into their agency to find solutions. “Like anything, don’t try to boil the ocean. You need to start off with something small,” said Hawley. By starting small and iterating, agencies can learn the best practices and strategies to drive organizational change.

“There's all these new technologies coming out that can take advantage of many machines and we can shift through billions and trillions of records in an unprecedented amount of time,” said Carroll. “And that’s where you start to get away from the cost, but you get to delivery of new capabilities, new types of analytics and new recommendations.”

Work with Experts

The enterprise data hub continues to be an emerging trend for government, and it requires partnership with experts. Moving away from the traditional relational database and siloed information is a cultural and technological shift for agencies.

“[Cloudera] has the original founders of Hadoop. Our whole organization is built around making this platform an open architecture that we can leverage,” said Hawley.

In today's world, data is the key to innovation, and through progressive data management architectures like the EDH, agencies can turn their data into insights, and transform the way they operate.

“Data should be at the heart of your decision-making,” said Mudge. By taking on this philosophy, organizations will continue to drive innovation and transform the business of government.
About Cloudera

Cloudera is revolutionizing enterprise data management with the first unified platform for big data: the enterprise data hub. Cloudera offers enterprises one place to store, process and analyze all their data, empowering them to extend the value of existing investments while enabling fundamental new ways to derive value from their data.

Founded in 2008, Cloudera was the first and is still today the leading provider and supporter of Hadoop for the enterprise. Cloudera also offers software for business critical data challenges including storage, access, management, analysis, security and search.

With over 15,000 individuals trained, Cloudera is a leading educator of data professionals, offering the industry’s broadest array of Hadoop training and certification programs. Cloudera works with over 700 hardware, software and services partners to meet customers’ big data goals. Leading organizations in every industry run Cloudera in production, including finance, telecommunications, retail, internet, utilities, oil and gas, healthcare, biopharmaceuticals, networking and media, plus top public sector organizations globally.

Learn more at: http://www.cloudera.com

About GovLoop

GovLoop’s mission is to “connect government to improve government.” We aim to inspire public sector professionals by serving as the knowledge network for government. GovLoop connects more than 100,000 members, fostering cross-government collaboration, solving common problems and advancing government careers. GovLoop is headquartered in Washington D.C. with a team of dedicated professionals who share a commitment to connect and improve government.

For more information about this report, please reach out to Pat Fiorenza, Senior Research Analyst, Gov-Loop, at pat@govloop.com.

1101 15th St NW, Suite 900
Washington, DC 20005
Phone: (202) 407-7421 | Fax: (202) 407-7501
www.govloop.com
Twitter: @GovLoop

Access and Use of NASA and Other Federal Earth Science Data

Source: http://commons.esipfed.org/node/2419
Abstract/Accord

This session will address the access and use of Earth science data that is offered by NASA and other federal agencies. Data users will describe their experiences in accessing and using data to analyze environmental phenomena or create mobile applications. The session will feature a panel of federal data users and a discussion with data users and providers.

Notes:

Access and use of NASA and other federal Earth science data

Lead: Ethan McMahon

Update: IT&I Rant & Rave will be Thursday, October 2, 3 ET and will follow up from the summer meeting panel on this topic.

Rant & Rave Title: Access and use of federal Earth science data

Are you curious about what data users want? How can we help new audiences for Earth science data, like the interested public and developers, answer scientific questions or create decision support tools? Call in to the Rant and Rave session on Thursday at 3 PM Eastern time to explore these questions in more depth.

At the ESIP summer meeting, a panel offered their views about the access and use of federal Earth science data (see http://commons.esipfed.org/node/2419 for notes and presentations). On Thursday we can follow up on that discussion and talk about these questions. Feel free to edit this page with your suggestions and please bring your experience and ideas to the discussion!

- What do new users of ES data (interested public and developers) want?
- How can we find out what they want?
- What audiences should the data provider community focus on?
- What can the data provider community do to help them?

Notes from the panel at the summer meeting

Session is focused from the data user perspective introduction, panelists describe experiences, followed by discussion.

There are many earth science data users. Researchers, interested public, software designers

Increasing interest from public and software developers. Preform their own analyses/citizen scientists. And those

making green apps or have desire to help but do not have the context.

Working with developers on apps, they posted problem statements and data. Then asked them to see what they could come up with, with open source code and systems. Two good outcomes, and many prototypes that can become more robust products. The developers have lots of questions, and that means waiting for emails or calls, and it is often not
what the producers have expertise or knowledge to answer. Also developers might not be successful because of poor documentation, not experts in this area, or unusable data.

Mentioned a few examples, see slides.

White house initiative - Big Earth Data Initiative. Common approaches to discoverability, accessibility, and usability. Hoping to figure out how people can do their work so we can get to the analysis stages.

Speakers:
Tamara Ledley of TERC
Janet Fredericks of Woods Hole Oceanographic Institution
Margaret Mooney of UW-Madison
Rob Carver the weather channel

Ethan introduced each speaker before their presentation.

**Earth Exploration Toolbook (EET)**

Tamara Ledley
Making geoscience data accessible and usable in education.

Each chapter provides step by step instructions for access, use and analysis for the data. It is for the teachers so they have enough experience to create an activity for their students. There are 44 chapters.

How do we get to creating these chapters? Once we have the format? Started with DLESE. Found a data provider and then bridge the gap between the creators and educators. Available in digital libraries.

Each team for data access - has 5-6 people. Data provider, tool specialist, scientists, curriculum developer, educator. These vary depending on the topic of the chapter. The areas of expertise were treated as peers with equal weight in the conversations. For many educators this was the first time they were treated at this level.

They held 6 access data workshops from 2004-2009. They had ESIP funding for one through the funding Friday events. There were 240 participants and 10 teams per workshop. They conducted an evaluation with primary and secondary roles. In the evaluation, they asked what people found most valuable. What was striking, was people wanted more team break out time. Also networking with those in other fields was also important. Educators and scientists or technologists.

They conducted a longitudinal study of the science/tech and educational communities. Meeting science standard and pattern recognition was higher in the educators.

Other results related to “what do you do with the data?”. Depended on educational community as well.

Longitudinal impacts survey - gave some examples from the results related to data specialists. Gave some examples of how the workshops impacted their process in managing data and making it available.
Review criteria for data sets - gave examples of some of these points. Ex: Curriculum developers can find and use data easily - for high school vs. undergrad etc.

Data sheets - as published in EOS in 2008, Ledley et al. Gave example of the educationally relevant metadata standards for data sets.

Data sheets website has more examples of collections etc. We looked at one of the websites http://serc.carleton.edu/usingdata/browse_sheets.html walked through selecting a data set and exploring data found.

EET utilization metrics - looking over 7.5 year range. Visit duration, page depth, returning vs. new visitors.

Experiences Accessing Federal Data

Janet Fredericks

http://www.whoi.edu/mvco

Talking about two stories - 1) data accessed to look at a coastal breach, and 2) another looking at satellite data - elevations etc.

1) went to google and looked for time breach information - when did the breach happen? Found it happened between July 2006-July 2007. Google did not help so went to Data.gov. Found the viewer she had used 3 years ago, updated browser, logged in, downloaded data needed, found April 1st and May 3rd as a more specific date range. Then went back to the data and was able to be more detailed/direct.

2) Looking at features on the Mid Atlantic continental shelf, and if the conditions were changing. Talked to people in the community. The links that worked in 2010 no longer work now when checking the sources. Then she looked at the coastal relief model. Tried a few others seafloor etc. but found the right one. In the past had to download the individual blocks of data, but now you could get the whole thing. In blocks this was hard to reconnect. Had to deal with two volumes and multiple blocks. But new features let her just put in coordinates. But finding the tool was difficult and hard to find again. In 2010 used Rich Signell’s query you could use in Matlab. Had to update many things when retrying in 2014, but used the service on his site. To get actual data instead of a map.

Confessions of a data Hoarder

Rob Carver

Open data and the Weather Company - took open data and becoming a big business. Usually from weather service, to create forecasts and tell stories. Over 100 TB of data in many locations. Model data, radar, shapefile data (FEMA, census bureau etc).

Locating data - google and literature search, ???, Data! Had to make evaluations about the project to pick sources that showed up in the results.

Most data accessed in unidata manager. and ECMWF pushes down data. This is ingested in to forecast system and GRADS. Then it is archived on local disk arrays and Amazon 3S.
NCSC archives - order data, pull down two year sets because it is easier than asking for a subset.

FEMA flood maps - problematic acquisition. DVD per state, large shapefiles. Had to parse them and break it down by county.

Suggestions - data in a difficult/proprietary format is just wasted disk space; use well supported open source software packages for data formats; instead of complex CSV files use self describing formats like JSON, XML, etc.; data/navigation files should use same naming process; don’t use overly large archive files, data pools attached to large disk arrays are awesome; for really large data sets bittorrent would be useful.

Applications of satellite data used in the SatCam and WxSat mobile apps

Margaret Mooney

SSEC data center has various data, which is funded by soft money and it is free. To push it out is a fee. Two apps, WxSat and SatCam. In SatCam there are 4 different screens. Walked through how the app works, with satellite images. Working to add more data on air quality to SatCam.

Discussion

Matt - How was visualization used while verifying that you have the data you want before you download in ascertaining data quality and or in the metadata standards? What worked and what more would have been wanted? Tamara said those were suggestions and not guidelines. To ask them how the workshops impacted and these were suggestions of things that would be needed.

Jeff from NOAA - question to Rob, working on ways to make things better, but specifically asked about GRIB a good or bad format? Rob - I am used to GRIP so maybe Stockholm Syndrome. Obvious GRIP one is bad but GRIP two get better. and it is open.

Question - given satellite data becoming bigger, should we push grabbing data vs central servers? Maybe ship algorithm and then we can send you the data rather than asking for all of it? So bandwidth is not tied up? Rob - that might have to be a community process in the cloud.

Follow up to that - does most of Carver’s data get processed by WSI? Follow up to follow up - does he see WSI be the baseline, will WSI put base stations down for others? Answer from audience - SMAP made decision to not make direct broadcast capability But other venues and such may Two issues - latency vs bandwidth

Answer to Ethan’s question to group - construct services that people can build on, that everyone can use.

Rama adds, you can’t have one set of services - because you have different types of users.

Follow up from Ge - should we make datasets catered to different users, or have datasets follow standards to then use services for the users?
From Rama: What about the value chain? Others adding to value chain will be on the backs of agencies.
Answer: First and foremost - original has value, get that metadata. Then move toward interoperable.

There is a way to have the data to be available, and then you might later offer them in other "easier" formats.
World moving towards data-services…

The DAACs get metrics from how many times their data sets are downloaded. Two questions for panel - How can we get more information that would be useful? What are the biggest stumbling point?

### Attachments/Presentations

<table>
<thead>
<tr>
<th>Attachment</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="#">ESIP data access session EMcMahon 7-8-14.pptx</a></td>
<td>330.42 KB</td>
</tr>
<tr>
<td><a href="#">Making Geoscience Data Accessible and Usable in Education 0714 v3.pptx</a></td>
<td>28.35 MB</td>
</tr>
<tr>
<td><a href="#">ESIP_JFredericks.pptx</a></td>
<td>15.51 MB</td>
</tr>
<tr>
<td><a href="#">WxSat&amp;SatCamJuly2014_Mooney.pptx</a></td>
<td>33.14 MB</td>
</tr>
<tr>
<td><a href="#">summerESIP2014_rcarver_powerpoint.pptx</a></td>
<td>1.15 MB</td>
</tr>
</tbody>
</table>

### Details

Tuesday, July 8, 2014 - 14:00 to 15:30
**Summer Meeting 2014**
Session Type: **Breakout**
Room Location: **Ptarmigan A**
Expertise Level: **Intermediate**
Accepted: **Yes**

### People

Session Leads:
Ethan McMahon

Presenters:
Participants

The scheduled panelists include: Tamara Ledley of TERC will describe climate and related data that is used by the Climate Literacy and Energy Awareness Network. Janet Fredericks of Woods Hole Oceanographic Institution will discuss data related to oceans. Margaret Mooney from the UW-Madison Cooperative Institute for Meteorological Satellite Studies will cover applications of satellite data used in the SatCam and WxSat mobile apps. Rob Carver of The Weather Channel will cover access of high volume weather data. Audience attendance list: Denise Hills, Stacie Bennett, Gee Peng, Steve Olding, Laura Ellison, Mary Armstrong, Ed Johnson, David Golaher, Christine White, Jeff de la Beaujardiere, David Bassendine, Rama, Jim Barrett, Kevin Browne, Janet Fredericks, Tyler Stevens, Tamara Ledley, Karen Moe, Gao Chen, Joe Roberts, Brian Blanton, Josh Young, Steve Kempler, Rob Carver, Cherly Levay, John Kusterer. Online participants: Gary Randolph, Jennifer Davis, Matt Cechini.

Citation

McMahon, E.; Access and use of NASA and other federal Earth science data; Summer Meeting 2014. ESIP Commons, April 2014
Submitted by erinmr on 04/22/2014 - 15:44.

Big Data in Materials Research and Development

Source: http://www.nap.edu/openbook.php?record_id=18760 (PDF)

Overview

Authors

Maureen Mellody, Rapporteur; Defense Materials Manufacturing and Infrastructure Standing Committee; National Materials and Manufacturing Board; Division on Engineering and Physical Sciences; National Research Council

Description

Big Data in Materials Research and Development is the summary of a workshop convened by the National Research Council Standing Committee on Defense Materials Manufacturing and Infrastructure in February 2014 to discuss the impact of big data on materials and manufacturing. The materials science community would benefit from appropriate access to data and metadata for materials development, processing, application development, and application life cycles. Currently, that access does not appear to be sufficiently widespread, and many workshop participants captured the constraints and identified potential improvements to enable broader access to materials and manufacturing data and metadata. This report discusses issues in defense materials, manufacturing and infrastructure, including data ownership and access; collaboration and exploitation of big data's capabilities; and maintenance of data.
Big Data in Materials Research and Development
Summary of a Workshop
Maureen Mellody, Rapporteur
Defense Materials Manufacturing and Infrastructure
Standing Committee
Division on Engineering and Physical Sciences

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, N.W. Washington, DC 20001
NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine.

This study was supported by Contract/Grant No. W911NF-11-C-0212 between the National Academy of Sciences and the Department of Defense. Any opinions, findings, or conclusions expressed in this publication are those of the author(s) and do not necessarily reflect the views of the organizations or agencies that provided support for the project.

International Standard Book Number 0-309-0XXXX-X
Library of Congress Catalog Card Number 97-XXXXX

Cover image: Big Data is partially driven by today’s ever increasing ability to electronically generate and store data. Big Data is also requiring ever increasing computational power to fully mine and understand the data. Thus, beside Big Data itself, at the core of this topic is today’s materials research and development enabling the bootstrapping of the complex electronic materials needed for the computations.

Artist : Erik Svedberg, image created electronically and algorithmically.
Acknowledgement of Reviewers

This report has been reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the National Research Council’s Report Review Committee. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making its published report as sound as possible and to ensure that the report meets institutional standards for objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process. We wish to thank the following individuals for their review of this report:

David Aspnes (NAS), North Carolina State University,

Valerie Browning, ValTech Solutions, LLC,

Jesus de la Garza, Virginia Polytechnic Institute and State University, and

Sylvia Johnson, NASA Ames Research Center.

Although the reviewers listed above have provided many constructive comments and suggestions, they were not asked to endorse the views presented at the workshop, nor did they see the final draft of the workshop summary before its release. The review of this workshop summary was overseen by David W. Johnson, Journal of the American Ceramic Society. Appointed by the NRC, he was responsible for making certain that an independent examination of this workshop summary was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of this summary rests entirely with the author and the institution.

Overview

The Standing Committee on Defense Materials Manufacturing and Infrastructure (DMMI) convened a workshop on February 5 and 6, 2014, to discuss the impact of big data on materials and manufacturing. The DMMI standing committee is organized under the auspices of the National Materials and Manufacturing Board of the National Research Council (NRC) and with the sponsorship of Reliance 21, a Department of Defense (DOD) group of professionals that was established in the DOD science and technology community to increase awareness of DOD science and technology activities and increase coordination among the DOD services, components, and agencies.

This report has been prepared by the workshop rapporteur as a factual summary of what occurred at the workshop. The planning committee’s role was limited to planning and convening the workshop. The views contained in the report are those of individual workshop participants and do not necessarily represent the views of the workshop participants as a whole, the planning committee, or the National Research Council.
To conduct its workshop on big data, the DMMI standing committee first organized a workshop planning committee to identify workshop topics and agenda items, speakers, and guests to be invited. The planning committee consulted with Reliance 21 and members of the community to develop and organize the workshop. The workshop was held at the National Academy of Sciences building, 2101 Constitution Ave., NW, Washington, D.C. Approximately 50 participants, including speakers, members of the DMMI standing committee, Reliance 21, invited guests, and members of the public, participated in the two-day workshop.

Some of the topics addressed at the workshop included these:

- Any unusual aspects of materials and manufacturing needs compared to other big data efforts.
- Data ownership and access, including materials property data.
- Collaboration and the exploitation of big data’s capabilities, including information exchange, validation, and security.
- Cost and ease of maintenance of data, including any associated infrastructure needed.
- Assuring the pedigree of the data and metadata.
- A general understanding of the materials community’s wants and needs in big data.

To assist the reader, a short summary of recurring themes from the workshop presentations and discussions is presented below. These six themes are merely a short description of the items that were discussed by multiple speakers or participants during the course of the workshop. They were identified for this report by the rapporteur, not by the workshop participants.

1. Data availability.
2. Data size: “big data” vs. data.
3. Quality and veracity of data and models.
4. Data and metadata ontology and formats.
5. Metadata and model availability.
6. Culture.

Within each theme, the discussion addressed current concerns, needs, or requirements—in other words, “challenges”—and potential improvements.

After briefly describing the recurring themes, the report summarizes the workshop presentations and discussions. Appendix A contains the statement of task for the workshop. Appendix B lists the workshop participants. Appendix C is the workshop agenda, and Appendix D defines acronyms used in this report.

Workshop Themes

Much of the workshop discussion was driven by an overarching assumption: The materials science community would benefit from appropriate access to data and metadata for materials development, processing, application development, and application life cycles. Currently, that access does not appear to be sufficiently widespread, and many participants
captured the constraints and identified potential improvements to enable broader access to materials and manufacturing data and metadata.

**DATA AVAILABILITY**

Data availability was a much discussed topic at the workshop. Several participants spoke of the difficulties associated with the fact that data are not always archived properly for long-term storage and that experimental data, including information about the procedures used and how the data were acquired, are not readily available in the public domain. Among the obstacles to data sharing discussed at the workshop were these:

- **Intellectual property constraints.** Several participants said that data ownership is an obstacle to data sharing. Thom Mason, of Oak Ridge National Laboratory, said that because individual researchers in the materials research community invest substantial time and effort in making and characterizing a sample, there will be challenges associated with moving into open source data. Other participants noted that access to proprietary and other data owned by private industry can also be constrained because a company wishes to retain its own intellectual property. This then becomes a cultural consideration in the community (see the final theme in this chapter, culture).

In discussion sessions, some participants also identified several challenges when the data are generated under a DOD contract. DOD contracts have many data requirements in them, and the DOD needs to be aware that these requirements can sometimes be perceived as cost prohibitive. One participant pointed out that material suppliers, original equipment manufacturers, and the government have a joint responsibility to share data. In some cases, a participant noted, suppliers provide the material but no corresponding metadata. Agreements with suppliers can take a year or two to develop, which can slow the rate of innovation. Julie Christodoulou, of the Office of Naval Research (ONR), stated that the Lightweight and Modern Metals Manufacturing Innovation Institute (LM3II) is currently sorting through issues related to levels of engagement and intellectual property. She believes that each project will likely have its own unique intellectual property arrangements. Other participants indicated that the Air Force Office of Scientific Research (AFOSR), ONR, the Defense Advanced Research Projects Agency (DARPA), and other defense agencies are exploring ways to share with the broader community data generated under a DOD contract.

- **Data heterogeneity.** Several participants pointed out that data are produced by many different people and organizations. They noted that when data are produced via heterogeneous distributed systems, there is no easy or centralized access.

Several possible remedies for the unavailability of data were explored at the workshop:

- **Data-mining utilities.** A participant noted the possibility of using data-mining utilities—something akin to a “materials Google”—to search existing data. This could improve access and increase ease of use.

- Storing materials instead of data. On several occasions, participants discussed the option of storing samples of a material directly instead of storing data related to an experiment. During a discussion session, Denise Swink, a private consultant, asked if reestablishing critical material repositories in the DOD should be considered. She said there is a reluctance to do so but noted the large financial difference between stockpiling and archiving—a few of the workshop participants are probably more interested in archiving. One participant mentioned the Air Force’s digital twin program, in which the digital representation of a material keeps information about the material properties. Perhaps it would be valuable to also save an actual sample to examine along with the digital twin. However, some workshop participants pointed out that critical questions would need to be addressed, such as how much material to retain, access criteria, and other policy issues. Others suggested that any sort of repository could be prohibitively expensive: To be a viable option, it might be necessary, they suggested, to set priorities for which materials to retain.

- **Government mandate to store data.** Several participants suggested that a data storage mandate might be useful, proposing that any government-funded data be put into a standard format and stored in some long-term external
A participant reported that this step is under consideration by the National Science Foundation (NSF) and the Department of Energy (DOE). A DOD participant said that DOD had explored the idea of such a mandate but found it to be time consuming and expensive and, in the end, not cost-effective. He said that a mandate would have to be for more than a data format; it would need to answer questions such as who owns and maintains the data and where the information should reside. Someone else noted that NSF has begun developing a data repository that it not, however, very user friendly.

- Development of private data repositories and formats. Workshop participants also discussed data repositories and how to ensure that data remain accessible in the future. Several persons spoke of the need to define structured databases. Dan Crichton, of NASA’s Jet Propulsion Laboratory (JPL), said that data archives typically require at least a 50-year expected time window for operability. For data to remain usable for that long, he pointed out, they would need to be captured in a stable format, even if that format is not contemporary, and should be software independent, relying upon static conceptual models. He noted in particular the importance of separating the data from the technology used to acquire it; otherwise, the data could be rendered obsolete too rapidly, as technology changes so rapidly. Michael Stebbins, of the White House Office of Science and Technology Policy (OSTP), discussed at length the White House open data initiative, which leverages existing and emerging information channels (such as journal publications) to create data repositories of the future. Another participant suggested that the DOD assess the cost of storage vs. reproduction of the data to determine if some sort of data repository (whether mandated or voluntary) makes sense. Yet another participant pointed out that it is fairly common for universities to have permanent storage facilities available, using the Deep Blue program at the University of Michigan as an example. However, other participants pointed out that these programs are expensive for the host and do not always include metadata.

- Automatic data capture. Several participants briefly mentioned the use of automatic methods to capture raw and unstructured data from journal articles or to capture and upload raw data from instruments during an experiment. Dave Shepherd, of DHS, discussed an automatic data capturing project in biosecurity known as algorithms for analysis, an algorithm-oriented project to identify emerging technologies that could be used against our nation. It uses natural-language processing software to find descriptors in the scientific literature, patents, and other scientific documentation. The data must be up-to-date and continually harvested. Such a project might be useful for the materials community as well.

DATA SIZE: “BIG DATA” VS. DATA

Several speakers noted the recent overall rise in scientific data output. For instance, Mr. Crichton showed that the amount of data produced at JPL has been increasing at a significant and highly nonlinear rate: in 2000, a planetary mission data set contained 4 TB of data. Now, such a project consists of over 500 TB of data. During their presentations, Jed Pitera, of IBM, Mr. Shepherd, and Chuck Ward, of the U.S. Air Force, all noted huge increases in data rates in their communities as well.

However, multiple participants repeatedly questioned whether materials science truly has a “big data” problem or whether it has simply a “data” problem. Several participants noted that the growth in data in materials science, while substantial, is not as dramatic as it is in other scientific disciplines. Dr. Swink argued that important data issues in materials science, such as proprietary data access and the lack of homogeneity, are data problems but not “big data” problems. Other participants indicated that, regardless of the regime, the amount of data produced in materials science is outpacing the algorithms and processing needed to analyze it. They pointed out that the traditional model of data analysis, which relies on the individual researcher, is not likely to be a viable model as the data size continues to increase. At the local level, there may be more data than the researcher has time to analyze. This could indicate need for more centralized analysis and computing tools. Several participants suggested looking at successful big data analysis techniques used in other domains, such as protein databases and work flow models.
Dr. Pitera suggested that the materials science community may be well served by using data reduction or extraction techniques so as to exit the big data regime—in other words, to make the “haystack” smaller and the search for the needle easier! (The idiom “to find a needle in a haystack” was taken up by participants and repeated many times over the course of the workshop.) However, other participants wondered how to perform data selection in the most judicious and domain-specific manner. One person mentioned the need to determine the necessary approach and its associated trade-offs: that is, whether one must interface with a large mass of data or whether one needs algorithms to select the salient features of the data one needs. In a discussion session, Jesus de la Garza, of Virginia Tech, pointed out trade-offs between collecting all the data possible (even if we don’t know what to do with it) and collecting only the information we know we want. He observed, however, that most research decisions require an assessment of trade-offs, and this is no different.

**QUALITY AND VERACITY OF DATA AND MODELS**

Many of the participants referred to the “four V’s” of big data (volume, velocity, variety, and veracity) several times during the course of the workshop. Although no one ranked them in importance with respect to materials science, much discussion time was devoted to the last two—variety and veracity. Individual participants noted that there are many sources of uncertainty in materials data, with no consistent methods to verify data quality. This problem underscores the benefit of openly sharing data, so that they can be verified independently.

Dr. Pitera pointed out that data quality can be limited by instrument quality, as well as by interpretation quality, and it can be influenced by various data artifacts. Mr. Crichton argued that peer review of data is necessary to assess usability; the international community should agree to standard models and to a consistent peer review process. He suggested that research supported by taxpayer dollars should make its data available and citable: Citations would give the reader confidence that the data set is publication-quality. Several participants discussed challenges associated with the verification of proprietary data. Someone noted that journal publishers are a good point of leverage in the academic community; however, in disciplines such as pharmaceuticals and some others, publication is not a high priority. The data in those instances are used to build business and are considered proprietary. It becomes difficult, if not impossible, for the outside community to gauge the quality of those data.

Several workshop participants complained that, in the material science community, it is difficult to compare the results of simulations to physical measurements. To help resolve this, Jesse Margiotta, of DARPA, noted that Integrated Computational Materials for Engineering (ICME) has a vigorous verification and validation component used to provide confidence limits. Another participant suggested using a materials work flow platform (such as Kepler) to capture data and enable reproducible results.

**DATA AND METADATA ONTOLOGY AND FORMATS**

Another challenge lamented at the workshop was the lack of standards and terminologies for data and metadata. Several participants noted the absence of a formal ontology for materials science and the need for a practical set of identifiers and descriptors. It was noted during a discussion that the materials community may suffer from a dearth of conversation about ontology.

One participant complained that the process for developing standard terminology is very difficult and slow, pointing to the erstwhile ASTM International committee that once developed standards in this area. In general, companies do not like to pay their employees to do this type of activity, however, and the ASTM International committee folded because...
there was no funding for it from the community. The same participant noted that companies may not be interested in attaching themselves to a certain standard format, because they are concerned they will be forced to share information they would prefer to keep proprietary in their own formats. During his presentation, Mr. Crichton suggested that the international community should agree to standard models and to a consistent peer review process. He pointed out that agreement on how to represent data can be very difficult because different scientists will have different emphases within the same data set. Mr. Shepherd also pointed out problems with many formats, resolutions, and source locations of large data sets.

To move forward, one participant suggested looking to the NSF program EarthCube 3 as a model for working across different communities to develop ontologies and names. Dr. Margiotta also reported that DARPA, along with the Army Research Laboratory (ARL) and other program partners, is developing methods to standardize data fields and metadata fields for materials and materials processing.

### METADATA AND MODEL AVAILABILITY

The concept of metadata availability had several meanings at the workshop. Mainly it referred to access to knowledge and information about a particular experiment—models used, starting conditions, and other “meta” information about the data—in other words, information that is not generally available in a journal article, which limits one’s ability to replicate the experiment. Many participants spoke over and over about the need to capture and report metadata.

In other cases, metadata availability referred to broader access to the models themselves; there were several separate discussions about the need to have a standard modeling toolkit available to researchers. A participant suggested that, for the small-scale researcher, data production is outpacing computing capabilities. In the case of neutron beams, for example, the bottleneck is in processing data, not in having access to beam lines. Continued progress is needed in developing new data analysis programs. Several participants and speakers discussed the idea of transitioning data computation to the source of the data. (This is also related to the workshop discussions of data size, summarized above.) Mr. Crichton said that this paradigm is likely to become increasingly important in the next 5 years, as data sets become more and more massive. More analysis tools and capabilities will be needed at the site of the data repositories. Mr. Crichton predicted that merely distributing data would soon become an obsolete approach; instead, research services and analytics will be a more advantageous approach, as users will need services rather than just the data.

Mr. Shepherd, of DHS, described a project in predictive biology that uses knowledge-based (KBase) data to combine data for microbes, microbial communities, and plants into a single integrated data model. Users can upload their own data and build predictive models, which represents more of a community effort in big data. Its plug-in architecture will allow other laboratories to use their own algorithms to analyze the data. This project mixes big data, high-performance computing, and cloud architecture. This type of community modeling effort may be a useful example for the materials science community.

Several participants stressed the importance of developing and distributing models for predictive purposes, particularly to determine inspection and maintenance intervals and life prediction as a function of use. Dr. de la Garza stated the importance of prediction rather than reaction; he suggested that big data analytics for prediction would be a powerful tool in the materials community. Mr. Shepherd and Dr. Pitera both provided examples of predictive modeling in different areas (biosecurity and equipment maintenance, respectively).
Several participants noted that the materials community lacks a data-sharing culture. This was likely the result of a variety of factors, they believed. The most prominent factor was the reluctance, particularly on the part of private industry, to share proprietary data or other data that may have value. One participant noted that withholding materials data is considered, in order to gain a competitive edge, and companies want to protect their intellectual property.

Another factor discussed was the difficulty associated with working under sometimes complex DOD contract rules and regulations.

Several participants discussed the need for incentives for sharing data. They proposed specific incentive models for sharing and structuring materials data, in lieu of purchasing data ownership rights. Such incentives included the use of data citations, the development of quid pro quo relationships, and the creation of precompetitive partnerships. Some participants suggested using existing information channels to increase the opportunity for data sharing, such as linkages from journals to data repositories, consistent with the goals of the White House initiative on openess and access to scientific data.

Session 1: Introduction to Big Data

Session 1 of the workshop focused on introducing the concepts of big data and its application to several different disciplines. Presentations were made by Michael Crichton, Jet Propulsion Laboratory (JPL), Jed Pitera, IBM Research-Almaden, and Dave Shepherd, Department of Homeland Security (DHS).

FRONTIERS IN MASSIVE DATA ANALYSIS AND THEIR IMPLEMENTATION

Daniel Crichton, Director, Center for Data Science and Technology, Jet Propulsion Laboratory

Mr. Crichton explained that he manages multiple projects and programs in scientific data systems for planetary, Earth, and other disciplines. JPL has a data science and technology center that supplies the infrastructure needs and end-to-end systems for scientific data systems across multiple disciplines. Mr. Crichton said that he was also a member of the committee that wrote the NRC report Frontiers in Massive Data Analysis (2013) and would describe some of the findings and recommendations of that study.

Mr. Crichton stated at the outset that all observational science platforms, whether from a Mars rover, an Earth observing system, or a neonatal sensor system for an infant, have certain elements in common: All simultaneously provide numerous forms of measurement and observation. He noted that there are several challenges in space systems, and probably in other domains as well:

- Space systems are developed and deployed worldwide, and data are generated across complex, interconnected systems. Mr. Crichton said that there are many producers of the data; as a result, the data are highly distributed, with limited data sharing.
- Systems are heterogeneous and located in different physical systems. As a result, the data are stored in a variety of formats, and access to data can be difficult.
- The data sets are so massive that they stress traditional analysis approaches.

Mr. Crichton explained that the elements to collect data are traditionally built independently, with each element supporting a single type of data and data collection. The concept of big data is the opposite: Big data seek to look at the
end-to-end architecture needed to support an entire system. The big data approach considers ways to scale and systematize analyses. He showed that the amount of data produced at JPL has been increasing at a significant and highly nonlinear rate.

Mr. Crichton explained the importance of an architectural approach to big data. He said that the National Aeronautics and Space Administration (NASA), for instance, had captured its data in high-quality archives, but NASA did not originally consider the full data life cycle; instead, it pushed the burden of data analysis to the individual researchers. In contrast, the big data approach supplies a technical infrastructure with an overall architecture. In addition, Mr. Crichton stressed that the technical infrastructure should be built independently of the data and their structure. In general, technology changes rapidly, and data should not become obsolete on the same rapid time scale as the technology. Once the infrastructure is in place to support data capture, it is also important to be able to analyze it. Mr. Crichton suggested the adoption of more advanced techniques for data analysis to increase the efficiency of data analysis in a distributed environment. He gave the example of a space mission with an interferometer. In the past, such a mission would provide a few hundred terabytes of data; now that same mission format will provide tens to hundreds of petabytes of data. This data explosion requires changes to data storage, processing, and management.

Mr. Crichton briefly described the big data life cycle, from data generation to archiving and analysis. In between, data triage is conducted; if the data are too massive, data reduction steps may be necessary to reduce their overall size. However, any data reduction requires inferences to be developed about the data, adding uncertainty to the data.

Mr. Crichton pointed out that NASA has specific challenges to the standard data collection paradigm. He defined three challenges that are encountered in space data systems in particular:

1. How do we store, capture, and transmit data from extreme environments?
2. How do we triage massive data for archiving?
3. How can we use advanced data science methods to systematically derive scientific inferences from massive, distributed science measurements and models?

Mr. Crichton said that NASA’s data are collected by onboard instruments, and NASA is limited as to how much data can be returned to the ground. It is therefore interested in onboard computing to perform data reduction. However, the science community is not always in favor of data reduction. In any case, onboard data collection systems are likely to reach a capacity constraint in the near future, which will force a change in system implementation.

Mr. Crichton then listed a set of technology trends identified by the NRC report on massive data (NRC, 2013):

- **Distributed systems.** This trend includes different access mechanisms and ownership rules, data federation, linking, etc.
- **Scalable infrastructures and technologies to optimize computing and data-intensive applications.** Mr. Crichton noted that the trend is moving toward data-intensive computing rather than high-performance computing.
- **Service-oriented architectures.**
- **Ontologies, models and information representation.** Dr Crichton stated a need to agree on how to represent data; this can be very difficult, as different scientists will have different emphases within the same data set.
- **Scalable data-base systems with different underlying models.**
• **Federated data security mechanisms.** Different access permission rules will create both authentication and authorization issues.

• **Technologies for the movement of large data sets.** Mr. Crichton pointed out that this is a big challenge in the high-energy physics community in particular.

Mr. Crichton then described the framework for data-intensive systems built at NASA’s JPL. The framework is open source, and it has been applied in various settings by the broader community. He said that there is no single big data solution; instead, one must bring the different building blocks together for each problem in such a way that they can be scaled and optimized. The framework is known as Object Oriented Data Technology (OODT). OODT was originally developed to support NASA’s planetary data system, a system recommended by an earlier NRC study. The data were stored on tapes, and the quality was beginning to erode; an NRC study recommended the development of the Planetary Data System (NRC, 1995). The purpose of the system is to collect, archive, and make accessible digital data and documentation produced from NASA’s exploration of the solar system. Mr. Crichton explained that the data have international operability and have been validated against a common set of structures and data standards.

Mr. Crichton then described the earth science data pipeline: Systems are built to capture data. Those data are sent to the ground station; higher-order data products are developed; the results are archived; and the results are provided to the community at large. This system has worked well; currently, it consists of a well-curated repository of about 7 PB, though that number will increase rapidly. Mr. Crichton said that JPL is also supporting the Earth System Grid Federation (ESGF), which informs the entire earth science modeling community and also supports the Intergovernmental Panel on Climate Change (IPCC) assessments.

Mr. Crichton then said it is important to evolve from data archiving to scalable data analytics. Data analytics requires automated methods to identify and detect data patterns across a variety of disciplines and under many different operational paradigms. He pointed out that a system could be used to study Twitter patterns for security purposes, for example. Mr. Crichton said that JPL is focusing on the transition from the data computation to the source of the data. This paradigm is likely to become increasingly important in the next 5 years, as data sets become more and more massive. More analysis tools and capabilities will be needed at the site of the data repositories. Mr. Crichton predicted that merely distributing data would soon become an obsolete approach. Instead, research services and analytics will be a more advantageous approach, as users will need services rather than just the data.

Mr. Crichton explained the elements to enable big data analytics for NASA. A schematic of these elements is shown in Figure 1. Currently, there are a variety of data sources, including satellite archives, airborne instruments, in situ data, and ground sources. These constitute a massive data set that has been brought together but which was not originally designed to be integrated. The data science infrastructure, such as the data, algorithms, and machines, informs data analytics. Big data triage includes the following elements:

• **Detection.** Identify elements of interest.

• **Classification.** Organize data automatically and in real time.

• **Prioritization.** Use the information to inform adaptive data compression.

• **Understanding.** Explain events for humans to understand and interpret the results.
Mr. Crichton pointed out that the same challenges exist in other fields, not just planetary science or earth science. For example, cancer research databases will probably require similar automatic processing of data, and astronomy and planetary science will also likely need automatic processing and feature detection.

Mr. Crichton then explained that the 2013 NRC Frontiers report emphasized how to systematize and automatically analyze data. The report also stressed the need to bring together experts in multiple disciplines to solve these big data problems.

Mr. Crichton concluded with several of his own recommendations:

- Invest in capturing and maintaining data in well-annotated, accessible, structured data repositories.
- Bring together cross-disciplinary groups to understand how to systematize massive data analysis and increase efficiencies.
- Develop computing architectures for sharing and analyzing highly distributed, heterogeneous data, aided by international cooperation and coordination.
- Enhance sustainability in both the data and the technology infrastructures while keeping the two on separate paths to maintain the long-term health of the system.

In the discussion period, several participants suggested mapping these ideas onto advanced materials and manufacturing. Some other participants noted that the materials and manufacturing communities may not even have the models to analyze the data; in this case, care has to be taken not to “throw the baby out with the bath water” during the data triage process. In the materials community, it is currently very important to capture the raw data. A few participants discussed the need to (1) understand the best way to downselect the data rather than to capture all of it and (2) focus on models to best understand the salient features of the data that are necessary to retain.

One participant asked about data repositories and how to ensure that data remain accessible in the future. Mr. Crichton responded that it is important to separate the data from the technology. Data need to be captured in formats that are stable, even if that means that the format is not contemporary. The data need to be software independent and to rely
upon conceptual models that are static. Mr. Crichton said that archives typically require at least a 50-year expected time window.

Mr. Crichton was also asked about how to address differing data quality. He responded that peer review of data is necessary to assess usability, and that the international community needs to agree to standard models and to a consistent peer review process. He said that federal research should make its data available and citable: Citations allow the reader to have confidence that the data set is publication quality.

**IBM AND BIG DATA**

Jed Pitera, Manager, Computational Chemistry and Materials Science, IBM Research-Almaden

Dr. Pitera explained that he manages IBM’s research team in computational chemistry and materials science at IBM Research-Almaden. He said that this group generates large volumes of structured data and that it focuses on computation more than on data. His research group examines computational problems in soft condensed matter (i.e., polymeric materials), which includes the computational screening of chemical reactions, the development of multiresolution algorithms, models for future technologies for lithography and nanopatterning, and models of fundamental interactions in chemistry and physics. In addition to his group’s data needs, Dr. Pitera pointed out that IBM also must analyze large amounts of unstructured data to support its business applications. To add nuance to the word “big” in big data, Dr. Pitera described some of its specific challenges. He stated that big data are often described in terms of the four V’s:

1. **Volume.** Volume refers to the physical amount of data. Dr. Pitera stated that problems related to volume in big data are scaling and the cost of dealing with that much information as much as structural problems.

2. **Velocity.** Velocity is both the rate at which data are collected and the time between receiving input and requiring an output. Dr. Pitera gave several examples of different time frames: In some instances, the maximum duration is several minutes (such as the time scale of a “coffee break,” the time it takes for a researcher to get a coffee and return to work); in other instances, the maximum duration is a few seconds (such as the time a call center operator might need to return an answer to a client).

3. **Variety.** Data come in a wide variety of formats.

4. **Veracity.** There are many sources of uncertainty associated with data quality. Data can be limited by instrument quality, as well as interpretation quality, and data can be influenced by various data artifacts.

Dr. Pitera stated that managing the four V’s requires expertise in data (acquisition, integration, cleansing, storage, protection, and management), domain (models and hypotheses), informatics (algorithms, simulations, and rules), mathematics (analysis), systems (scale and velocity), and visualizations (sharing results). He said that his group’s efforts focus primarily on the first two areas of expertise (data and domain): How do we get data, ingest them, and structure the domain of interest to capture the needed information? He posited that there are opportunities for big data in all types of study. The discovery phase focuses on conducting computational experiments, mining the literature (the published literature, as well as unpublished laboratory documentation), and finding new materials or repurposing existing ones. Dr. Pitera postulated that people tend to focus on the discovery phase. However, the material must be made usable, and
integration remains a challenge. Life-cycle issues, including continuous reengineering and monitoring, are all topics to be explored.

Dr. Pitera described three representative projects in materials and big data:

- **Harvard's Clean Energy Project.** This is a distributed computing project that focuses on using big data for materials discovery. It seeks to identify new photovoltaic materials with higher efficiency rates. This involves large-scale calculations, along with data mining and analytics. 5

- **Pharmaceuticals.** Pharmaceutical companies wish to mine patent data and medical literature to identify new relationships between drugs and disease.

- **Mining equipment.** The goal of this project is to understand predictive equipment maintenance for heavy mining machinery.

Dr. Pitera noted that the latter two projects are being developed in IBM’s Accelerated Discovery Laboratory, an exploratory collaboration space for experimental big data projects that include, government and academic and other private partners.

In closing, Dr. Pitera reminded participants that materials are complex, more than just a set of chemical reactions. Understanding a material requires knowledge about its statistical composition, processing history, and other metadata. He said that systems to manage data improve daily, and asked if an IBM Watson 6 for materials would be useful. He pointed out that if the same questions are asked repeatedly, then a Watson would be useful and the development costs could be justified. If not, what other type of infrastructure could be used? Finally, Dr. Pitera suggested that big data should perhaps be thought of as useful data that just happen to be big. He said that it may make sense to implement data reduction or extraction to leave the big data regime—in other words, it may be better to make the haystack smaller.

During the discussion period, a participant brought up a challenge unique to materials science: A material is qualified based on its characteristics, not its composition. Two materials may qualify for the same application but have different compositions. Examples such as aluminum alloys and jet fuel were discussed. Participants also wondered what the community should do next to come together. Dr. Pitera suggested that effort should be put into the construction of appropriate data models. These models are likely to be domain-specific, as metallurgy differs from polymer chemistry, which in turn differs from inorganic glass. A uniform description for data models would need to involve many different communities.

**BIG DATA FOR BIOSECURITY**

Dave Shepherd, Program Manager, Homeland Security Advanced Research Projects Agency, Department of Homeland Security

Mr. Shepherd began by noting that he works in biology programs with homeland security applications, and his portfolio does not include materials or manufacturing. He explained that the goal of biosecurity is to alleviate any accidental or intentional release of pathogens or other causes of disease. He said that biosecurity is a predictive field, seeking to prevent disease exposure and spread by means of anticipation, and it requires a long-term perspective. He said that this predictive need was identified by an OSTP workshop, predicting the Next Pandemic, in 2013. At the OSTP meeting, three of the five panel topics were related to big data (Big Data for Early Indications and Warnings, Big Data for Digital Surveillance/Digital Disease Detection, and Big Data Research and Analytics), and many of the problem sets discussed included the need for predictive tools.
Mr. Shepherd next explained that big data permit computer-assisted analytics, and algorithms and analytics can be used to classify and identify important information on a massive scale. Big data allow for correlations on a level that would not be possible with a smaller subset of information. One additional advantage for biosecurity is that surrogate data can be used. However, biosecurity big data sets tend not to be clean and do not contain consistent metadata. In addition, it can be difficult to find the right biosecurity data set for one’s needs.

Mr. Shepherd then went on to discuss three specific use cases for big data in biosecurity:

• **Use Case 1: Algorithms for Analysis.** This is an algorithm-oriented project to identify emerging technologies that can be used against the United States. It uses natural-language processing software to find descriptors in the scientific literature, patents, or other scientific documentation sources. The data must be up-to-date and continually harvested. Permissions are needed to access all the different possible data sources.

• **Use Case 2: Size and Architecture.** This project uses knowledge-based (KBase) data for modeling for predictive biology. KBase data combine data for microbes, microbial communities, and plants into a single integrated data model. These data are considered in a single, large-scale bioinformatics system. Users can upload their own data and build predictive models. This endeavor represents more of a community effort in big data. This project currently consists of 1 PB of stored data, and that amount is likely to increase significantly. Its plug-in architecture will allow other laboratories to use their own algorithms to analyze the data. This project mixes big data, high-performance computing, and cloud architecture.

• **Use Case 3: Surrogate Data.** Because the actual emergence of a novel biosecurity threat is rare, researchers seek to find surrogate data to help when new threats are encountered. Mr. Shepherd believed that the use of surrogate data is very helpful to this community, but that the idea of surrogate data may not be feasible in other communities. One example where excellent surrogate data for biosecurity exists, Mr. Shepherd noted, are data for the spread of antimicrobial resistance. Because outbreaks of antimicrobial resistance occur rather frequently, there are enough data to support predictive model development and verification. Like antimicrobial resistance, a novel biological threat can appear anywhere in the world, so infrastructure to support the development of a geographic distribution map would be useful and would assist in analysis and prediction.

Mr. Shepherd pointed out that clinical data are not aggregated at a national level, and there is no infrastructure to support such data aggregation. He said that researchers are beginning to realize that a centralized database is unlikely to be viable. When human subjects and privacy are involved, data sharing becomes difficult. Instead, it may make the most sense to move the data processing to the individual holders of clinical data. Mr. Shepherd envisions a new model in which clinical data holders participate in a national, distributed, interconnected grid similar to the collaborative model that underlies Lawrence Livermore National Laboratory’s Earth Systems Grid Federation for the climate modeling community.

Mr. Shepherd then described some of the challenges associated with big data in biosecurity, noting that these probably apply to other fields as well:

• **Data quality and access.** There are many formats, resolutions, and source locations. The use of images increases the data storage and transmission needs.

• **Personal data.** There are ongoing concerns related to personal data, privacy, and civil liberties.

• **Analytics.** Algorithms are not currently keeping pace with the enormous amount of data. More analytic processing capability is needed.

Mr. Shepherd concluded by saying that increased control over information has not led to increased biosecurity; instead, in the future, increased information sharing may be the way to increase biosecurity.
In the discussion, a participant suggested that it might be better to store the actual material instead of data describing the material. This brings up practical issues associated with how to store materials. This is true in biosecurity as well: Should you store the biological agent, data about the agent, or both?

A participant asked if the human immunodeficiency virus (HIV) and severe acute respiratory syndrome (SARS) would make good surrogates in biosecurity. Yes, said Mr. Shepherd, because emerging diseases, even ones that don’t represent a typical biosecurity scenario, can still provide modeling information.

Dr. McGrath pointed out that biosecurity is a large and difficult problem, larger in scope than what the materials community faces. He suggested interfacing with other communities that use predictive tools, such as the intelligence community. On the science side, he suggested working on sensor development and increasing basic knowledge and understanding.

A participant pointed out that cloud computing can be considered high-performance computing as well. Mr. Shepherd agreed and explained that he divided them in his discussion so that high-performance computing referred to anything from chips to run times, and that cloud computing referred to how to work together.

DISCUSSION

Ms. Swink opened the discussion session with a summary of several statements she had found compelling. One was the importance of determining the necessary approach and the associated trade-offs: whether one needs to interface with a large mass of data, or whether one needs algorithms to select the salient features of the data one needs. She also brought up the concept of ontology and asked whether this is a driver in the materials research and development area. Finally, she asked if the idea of surrogate data had any analog in the materials and manufacturing community.

Dr. de la Garza pointed out the need for a holistic perspective on big data, because they involve not only data collection but also life-cycle information: data collection, procurement, analytics, and stable, archival-quality formats. He also noted that there are trade-offs between collecting all the data possible (even if we don’t know what to do with it) and collecting only the information we know we want. He observed, however, that most research decisions require an assessment of trade-offs, and this is no different. Dr. de la Garza also underlined the importance of prediction over reaction; big data analytics for prediction would be a powerful tool.

Some participants also discussed the idea of storing materials vs. storing data. Ms. Swink asked the group to consider the reestablishment of critical material repositories within DOD. She indicated that she is aware of a reluctance to do so. She also argued that there is a large financial difference between having stockpiles or archives: The workshop participants were more likely to be interested in the less-expensive archives. One participant mentioned the Air Force’s digital twin program, in which the digital representation of a material keeps information about the material properties; perhaps it would be valuable to include an actual sample to examine along with the digital twin. However, other workshop participants also pointed out that critical policy issues would need to be addressed, such as the amount of material to retain, access criteria, and other issues. Others argued that any sort of repository, even an archive, might be prohibitively expensive. It may be necessary to set priorities on which materials to retain.

Participants also discussed common elements across disciplines that rely on big data. Some of these common challenges are associated with moving data, connecting to different ontology models to describe the data, and tools to identify compounds. Another popular idea across disciplines was bringing the processing to the data source and other
distributed data strategies. A participant noted that a Hadoop 7 might play a useful role in distributed data storage and processing, as part of a larger solution.

A participant suggested that the material science community look at examples in other disciplines to identify the costs associated with data storage. The participant suggested the fields of publishing, which is moving toward open access, and evolutionary biology, which has a significant amount of data but no repository structure. Another participant noted the utility of the open source model used by Dryad to provide access to data. 8

Ms. Swink asked what steps the community should take now to move forward in data management. One participant suggested looking to the NSF program EarthCube as a model for how to work across different communities to develop ontologies and names. 9 The materials community may suffer from the lack of conversation about ontologies.

Ms. Swink then asked if the materials science community has a “data” problem or a “big data” problem. She suggested that the problems are just data problems. She argued that important data issues in materials science, such as access to proprietary data and the lack of homogeneity, are not big data problems. She also pointed out that the long-time separation between materials development and product commercialization interferes with a company’s ability to understand the correlation and causality between product success and materials data.

Session 2: Big Data Issues in Materials Research and Development

Session 2 of the workshop focused more specifically on the application of big data to materials research and development. Jim Davis, University of California at Los Angeles, had originally been scheduled to present during Session 3, but the workshop participants determined his talk content was well suited to the discussions in Session 2, so he was moved here. Thom Mason, Oak Ridge National Laboratory; Chuck Ward, U.S. Air Force; Rusty Irving, General Electric; and Dr. Davis made presentations in Session 2.

PHYSICS IN BIG DATA

Thom Mason, Director, Oak Ridge National Laboratory

Dr. Mason began by noting that the title of his talk might be misleading: His presentation focuses on scientific data, he said, which includes physics, but it includes many other fields as well. He explained that Oak Ridge National Laboratory (ORNL) is a DOE laboratory, and its drivers are derived from the DOE mission. Many of these drivers involve large amounts of data and computing. He pointed out that the data requirements vary across scientific disciplines. High-energy physics and cosmology have very high data volumes, on the order of 1 PB per day. Data volumes for neutron sources are somewhat smaller, but they are still bigger than can be comfortably managed. Improvements in imaging systems have resulted in larger volumes of data, the scale of which can overwhelm a smaller research group with access to an ORNL beam line. Dr. Mason said that data come in many varieties (such imaging data, text, and sensor data), and the content must be combined. He argued that this growth in data is outpacing computing growth.

Dr. Mason also noted, as an aside, that a particular challenge in the materials community is that of provenance. Other disciplines, such as cosmology, have no concept of a sample. In materials, individual researchers invest much time and effort in producing and characterizing a sample, and as a consequence there will be challenges associated with moving to open source data because of the intellectual investment in a particular sample. The variability of samples between the different groups producing them is another challenge.
Dr. Mason stated that ORNL is DOE’s largest science and energy laboratory, and materials science permeates all of ORNL’s focus areas. He described ORNL’s high-performance computing ecosystem, including its flagship system, known as Titan. ORNL’s high-performance computing system models and simulates physical, biological, and chemical systems. It processes data, but it generates large volumes of it as well. The primary drivers for the supercomputing power at ORNL are scientific, but ORNL is also exploring industrial applications. Dr. Mason provided three examples:

- **General Motors.** General Motors sought to understand and develop new thermoelectric materials for higher fuel efficiency.
- **BMI.** BMI Corporation sought to create parts to retrofit onto Class 8 long-haul trucks for improved fuel efficiency and emissions.
- **Boeing.** Boeing sought to develop and validate computational models for transport airplane design and development.

Dr. Mason then described the Accelerating Data Acquisition, Reduction, and Analysis (ADARA) program at ORNL. ADARA is an experimental data analysis program developed in response to increased data and imaging needs from the Spallation Neutron Source. He pointed out that the Spallation Neutron Source is improving, and data volumes are growing at corresponding rates. In the past, the data volumes were handled via an individual researcher’s analysis code. The general rule of thumb was that one week of data collection equaled one year of graduate student analysis time. Now, however, the data are over a hundred times larger in size, and the old model of analysis has become unsustainable. When Dr. Mason was later asked if ADARA was limited to the Spallation Neutron Source, he clarified by saying that while ADARA is specific to the Spallation Source, the concept is applicable to other neutron and X-ray sources. The code is largely open source, facilitating expansion elsewhere.

Dr. Mason also described ORNL’s manufacturing demonstration facility (MDF), which includes extensive capabilities for additive manufacturing. ORNL is interested in expanding the palate of materials it uses for additive manufacturing. The laboratory is developing titanium alloys (primarily for the medical device community), high-temperature alloys, and refractory metals (such as tungsten). The processing recipes and real-world performance of these materials are still unknown. Dr. Mason pointed out that ORNL’s high-performance computing capabilities are useful to model the additive manufacturing of complex shapes. He then described several examples from the MDF:

- **Additive manufacturing of turbine blades.** ORNL and its industrial partner, Morris Technologies, use neutron scattering to measure atomic plane spacing and better understand the link between residual stress and laser additive manufacturing processing. They seek to develop a pedigree that will enable the turbine blades’ use even in situations that could otherwise compromise human or environmental safety.
- **Rapid, agile manufacturing.** ORNL and its industrial partner, Local Motors, use crowd sourcing to design vehicles for short-run manufacturing. The project focuses on novel material development and additive manufacturing. They are currently creating a process for large-scale (20 feet per side) carbon-reinforced polymer development that combines additive and subtractive techniques.

In the discussion period, Robert Schafrik asked about the additive manufacturing processes. He pointed out that the construction of a large structure takes time and will require changes to the additive process. Those changes might affect the properties of the end result and should be accounted for in the models. Dr. Mason pointed out that ORNL is moving toward parallel processing with large component sizes. One promising avenue to improve throughput is to use additive manufacturing for tool and die production, greatly speeding up the development cycle, while employing traditional high-throughput techniques for volume manufacturing.
The discussion briefly turned to access to data. Dr. Mason said that ORNL desires to move to an open source model. Another participant thought that embargo periods might be useful; access to data could be restricted for a period of time, then the data could be sent to a public repository. Another participant discussed the problems associated with a protein crystal structure databank; researchers are required to submit data to that databank. The community is discovering that much of the data is wrong, a sign that experimental constraints may need to be imposed for data consistency. Dr. Mason explained that, for the vast majority of work done at ORNL, the user owns the data but the expectation is that the data will be published. Under those constraints, ORNL does not collect a user fee. A small amount of work conducted at ORNL uses proprietary data, and that work is arranged under a full-cost recovery model.

MATERIALS GENOME INITIATIVE AND BIG DATA

Charles Ward, Lead for Integrated Computational Materials Science and Engineering, Air Force Research Laboratory

Dr. Ward began his presentation by showing that the amount of information related to materials has been increasing dramatically each year; in 2012, over 162,000 journal articles were published in the fields of materials science and engineering. He pointed out that industry has some of the best materials databases in the world. He gave an example of data-driven materials development in a cast and wrought disk alloy, R65, conducted by GE Aviation. Data-driven methods such as these are able to reduce both development time and costs by up to 50 percent.12

Dr. Ward then described the Materials Genome Initiative, which intends to create a Materials Innovation Infrastructure comprising accessible digital data, computational tools, and experimental tools. More about the Materials Genome Initiative can be found in Box 1.

Dr. Ward then defined data and metadata. Data are the result of measurement; data can be physical, consisting of experimental results and uncertainty, or virtual, consisting of simulation results and uncertainty. Metadata are the information that describes the measurement process; metadata can also be physical, consisting of the experimental setup and settings, or virtual, consisting of the explicit underlying model, simulation software, and input parameters.

Dr. Ward presented an example of a traditional approach to materials development that focused on measuring a diffusion coefficient. In the example, a researcher would measure the diffusion coefficient by tracking the root mean displacement of tracer particles, record the values, and publish the results (including metadata in the publication), typically represented as a diffusion coefficient. However, this reporting approach is dependent on a specific model (the diffusion equation). The paper assumes that the data fit the model, and it does not actually report the experimentally measured data.

Dr. Ward noted that the crystallography community has moved toward open data sharing; crystallography is one of several disciplines that are leading the way in open source data. The Crystallographic Open Database stores over 250,000 compounds and minerals. The International Union of Crystallography has mandated data archiving for its journals; as a result, one database (Crystmet) has over 150,000 entries of metallic and intermetallic structures, and another (Cambridge Structural Database, or CSD) has nearly 700,000 organic and metal-organic structures and almost 100,000 macromolecular structures available in its database. Dr. Ward suggested a next logical step for the materials community would be to capture data pertaining to CALPHAD.13 CALPHAD data also consist of metadata that increase in spatial and temporal complexity compared to crystallographic data.

Dr. Ward then posed a set of challenges in addressing materials data. He noted the following:
- *Materials have a pervasive application*. As a result, there is no single government funding agency that leads a cohesive materials research and development effort.

- *Materials have a nearly infinite scale of design variability*. There are few reference data sets, making it difficult to standardize data descriptions.

- *Materials are studied and used by many technical disciplines*. As a result, the data are widely dispersed, and a disparate vocabulary is used to describe them.

- *Materials are often a product differentiator*. This means that proprietary protections are often put into place for materials, making data sharing more difficult.

- *Materials are a key to economic and national security*. As result, they are subject to export controls, such as the International Traffic in Arms Regulation (ITAR) and Export Administration Regulations (EAR).

**BOX 1 Materials Genome Initiative**

The Materials Genome Initiative (MGI), announced by the White House in June 2011, aims to double the speed at which materials are discovered, developed, and manufactured. MGI seeks to develop a materials innovation infrastructure that includes the following:

- Computational tools for modeling, simulation, design, and exploration.
- Experimental tools for synthesis and processing, characterization and analysis, testing and prototyping, and verification and validation.
- Digital data.

The goal of the MGI is to develop collaborative networks that support the sharing of best practices and data to foster an open environment for materials design and development. Areas of impact include clean energy, human welfare, national security, and the next generation workforce.

**FIGURE B-1 Conceptual representation of the MGI**

showing the overlapping infrastructure requirements and the application areas.

Dr. Ward provided information on the National Institute for Standards and Technology (NIST) workshop on MGI data, held in May 2012. The workshop identified common themes to be addressed for materials data archiving, including the following:

- Materials schema/ontology.
- Data and metadata standards,
- Data repositories/archives,
- Data quality,
- Incentives for data sharing,
- Intellectual property, and
- Tools for finding data.

Dr. Ward also showed results from an open survey conducted by the Materials Research Society (MRS) and the Minerals, Metals and Materials Society (TMS) in the summer of 2013. The survey had a large number (675) of respondents. The respondents reported an interest in sharing fairly modest amounts of data overall (about half of the respondents said the data quantity they wish to share would be <1 GB per year). Survey participants also responded that they would be most interested in databases and data mining tools related to a material’s physical and thermal properties. This is actually the most basic level—physical and thermal properties have the least complex data and metadata requirements. Survey respondents noted some impediments to data sharing, particularly with respect to data ownership and intellectual property rules. One slight surprise was that receiving feedback from others was considered a strong motivating factor for sharing data, on a par with increased research visibility.

Dr. Ward then described MGI initiatives with data-intensive elements taking place within the government, including a number of projects at ARL, AFRL, ONR, DARPA, DOE, and NIST. He pointed out that these efforts use an ICME framework. (For more information about ICME, see Box 2.) The activities are geared to action. Dr. Ward postulated that the greatest challenges facing the government agencies in implementing MGI are the coordination of effort and the understanding of lessons learned by different agencies.
Dr. Ward stated that open access to research results through open access publishing is growing, and it is creating an entirely new model for publishing. The government is encouraging open access via data management plans required in proposals—both NIH and NSF have such requirements—as well as through a White House directive on public access (see Dr. Stebbins’s presentation summary, below, for more information). Some research disciplines, such as crystallography and evolutionary biology, have taken the initiative by adopting their own data archiving policies.

Dr. Ward concluded with the following summary statements:

• Materials data have intrinsic value that can enhance the research and development process.
• Several technical and cultural factors make the capture, archiving, and sharing of materials data difficult.
• The materials community appears willing to take on the challenge of materials data.
• Several MGI and community-led efforts are under way to guide materials data archiving.
• Broad discussion of materials data archiving within the community is gaining momentum and needs to be nurtured.

BOX 2 Integrated Computational Materials Engineering

ICME is “the integration of materials information, captured in computational tools, with engineering product performance analysis and manufacturing-process simulation” (NRC, Integrated Computational Materials Engineering, p. 9). It is a process by which materials, manufacturing processes, and component design can be optimized long before the components are fabricated. It is considered a promising emerging discipline that is still under development.

FIGURE B-2 ICME system that unifies materials information

into a holistic system that is linked by means of a software integration tool to a designer knowledge base containing tools and models from other engineering disciplines.


The question-and-answer period began with a discussion of data ownership. A participant asked about a scenario in which the data belong to whoever pays for it, not to the original data collector. Would this use of data without the knowledge of the persons who created the data not be dangerous? Dr. Ward pointed out that this is part of the normal
course of science: One person creates data, another refines it. He gave several examples, such as data related to genomics, in which 500 people might access and analyze the data without having been part of its collection. He also pointed out that the field of evolutionary biology saw a 70 percent increase in data citations after setting up the Dryad program.

A participant asked if these challenges were being addressed by the materials genome community, or if they were merely being listed in the hope that another community would address them. Dr. Ward responded that the path forward is fairly clear, and that MGI hopes to follow the model of the evolutionary biology community in bringing the community together to decide what to do and how to add value.

Another participant pointed out that NSF and the National Institutes of Health (NIH) require data management plans in their proposals, and universities now have systems in place to publish data sets and provide them with a digital object identifier, which is citable forever. These might be models for the materials community. A data citation index might also be useful.

**GENERAL ELECTRIC EFFORTS IN MATERIALS DATA: DEVELOPMENT OF THE ICME-NET**

Rusty Irving, Services Technology Leader, General Electric Global Research

Mr. Irving’s talk drew on information from the Metals Affordable Initiative Workshop on Data Management for ICME in June 2011. At that meeting, he listed four key advancements in computer science that should be taken advantage of:

1. **Cloud computing.** This is elastic, use-as-you-go computing.

2. **Service-oriented architecture.** This allows for interoperability and sharing.

3. **Federated data.** This allows the owner of data to share some parts and withhold other parts of databases. Users see data as if they were from one database.

4. **Business process management.** This is modern work flow software.

When combined, the four items can provide a collaborative ecosystem for material development. The outcome at GE is the ICME-Net, which, according to Mr. Irving, fosters collaboration, enables the reuse of common analytics and processes, and builds knowledge accumulation and sharing. The interface to ICME-Net is a web browser providing access to all in GE who are cleared for access. Mr. Irving provided an example of ICME-Net for disk forging, shown in Figure 2.

Mr. Irving explained that GE had teamed with MIT to develop a collaborative ecosystem for open design and manufacturing; MIT was able to provide the software components needed, and together they built an environment for crowd-sourcing military vehicle design. While the project did not move to Phase II, they were able to develop a robust prototype. GE leveraged the GE/MIT collaborative ecosystem platform technology to develop an innovation infrastructure for ICME-Net. The initial vision for ICME-Net consisted of the following goals:

- Enable geographically distributed collaboration on development, testing, and demonstration of ICME technology.
- Rapidly disseminate technology development through a “marketplace” of materials engineering models and processes.
• Enable the construction of large, complex simulations for ICME.
• Attract and sustain an ICME community.
• Provide an opportunity for both open source and retention of IP access to technology and best practices.

FIGURE 2 ICME-Net for a disk forging use case.


Mr. Irving then described ICME-Net in more detail. The goal is to grow a collaborative ecosystem. ICME-Net intends to build a marketplace that will subtly motivate the community to participate. This will enhance productivity, reduce development cost and time, curate model development, and make businesses more responsive. Projects consist of both components and services, and users can also be contributors.

Mr. Irving then discussed three use cases for ICME-Net. The first was the ceramic matrix composites project. ICME-Net provides a single location for the storage of experimental data and analysis. Users can decide whether to store a calculated result. The single location provides an auditable trail of every activity; it can be searched to find analysis done on a particular day or by a particular person. Mr. Irving said the interface has some similarities to Facebook, in that users can say whether they “like” a particular service. In the marketplace, therefore, the best services will accumulate the most “likes.” Mr. Irving explained that each element (users, concepts, components, subcomponents, and services) is kept small. Curation and metadata management can be incorporated as the project develops.

Next, Mr. Irving described the Materials Applications Engineer (MAE) for rotors, the second use case. This project results in large data sets, consisting of cut-up and microstructure data, as well as results from forging and heat-treatment models.

The final use case described was in alloy development. Users build small pieces of software to open the files and analyze the data one section at a time. The ICME-Net application contains a graphical user interface to lay out work flows and show linkages. The execution of the analysis is conducted in the cloud, and the curation and metadata management are separated from the data themselves.
Mr. Irving noted that ICME-Net can establish basic services for a particular application within several weeks. It was not originally intended to be used as a productivity tool, though that is now a top benefit. More functionality will continue to be added. He concluded with three main ideas:

1. The ICME concept is not new, and there are similar needs in other application areas, such as cyberinfrastructure.

2. Additional computing and software can extend the ICME-Net capability to, for example, a supply chain scenario.

3. There remain many outstanding issues associated with intellectual property, export control, proper business models, and other policy concerns.

In the discussion session, someone asked who the users of ICME-Net are. Mr. Irving responded that currently the users are all internal to GE, although GE is working to allow vendor access as well. The users are materials scientists with data that need to be analyzed.

SMART MANUFACTURING: ENTERPRISE RIGHT TIME, NETWORKED DATA, INFORMATION, AND ACTION

Jim Davis, Vice Provost, Information Technology, and Chief Technology Officer, University of California, Los Angeles

Dr. Davis pointed out that his talk would not focus on "big data" per se. He said his focus is manufacturing—but, as it happens, much is data-oriented in manufacturing right now. He began by defining some of the terms in the title of his talk. He explained that he means for the term "enterprise" to have as broad a scope as necessary, whether that means from factory to supply chain or across units. "Right time" is similar to real time, but it is associated with the window in time for taking action and not the rate of data collection. He stressed that "action" is an important element of smart manufacturing.

Dr. Davis then described the Smart Manufacturing Leadership Coalition (SMLC), which he cofounded in 2006. SMLC uses an industry-driven strategy to make more information available to the manufacturing community. The current focus of the SMLC is implementation. He explained that the manufacturing community generally wants to share services (or "apps"), not proprietary data. He said that in his role as UCLA’s chief technology officer, he interfaces with the manufacturing operations community, where he sees an emphasis on smart manufacturing, cyberphysical systems, and the Internet of Things (IoT). In his role as head of information technology for UCLA, Dr. Davis interfaces with the information technology community, where he sees a focus on enterprise resource planning, big data, cloud computing, and mobility.

Dr. Davis then provided several different examples of smart manufacturing systems and the role of data in those systems:

- **Example 1: Smart Manufacturing at General Mills.** This is an example of network-based manufacturing. Dr. Davis said that a "material" is a statistical composite of its constituent elements. In this example of General Mills, one of the "materials" is Cheerios. The material for Cheerios is supplied by farms, all of which are subject to varying conditions such as weather and water accessibility. It would be helpful for General Mills to have basic data from each of the many farms in its supply chain, such as lot size and other information about the lot. There are manufacturing constraints as well, as the production facility must comply with cleanliness and contamination.
standards, and the FDA has tracking and traceability requirements. Dr. Davis explained that General Mills has a “green light” procedure, in which the constituent elements are analyzed for readiness to be put into production; the recipe is confirmed; the material is made; and the product is released upon confirmation that it meets requirements. Dr. Davis noted that while the Cheerios application may not qualify as big data, a lot of supporting data are involved in the manufacture of a Cheerio.

- **Example 2: Smart Manufacturing at Praxair.** Praxair supports oil and gas production at 40 or so facilities worldwide. Its furnaces must be kept in a certain temperature range. Dr. Davis explained that Praxair is currently very conservative about that range, which leads to waste heat and extra associated expense. Praxair would like to manage its risk differently to realize the potential for significant cost savings. The furnaces cannot have sensors in them, as the environment is too harsh. Instead, Praxair has turned to high-fidelity modeling in production. This project involves data at many different scales.

- **Example 3: Smart Manufacturing at General Dynamics.** At its Scranton, Pennsylvania, plant, General Dynamics has activities in heating and forging as well as cutting and machining. It is attempting to match the requirements for heating and forging to the output of the cutting and machining operations to better manage its energy usage and production efficiency.

- **Example 4: ICME at Caltech’s Materials and Process Simulation Center.** This center assesses the manufacturing interface with materials models. The center is often asked about material risk and response and would like to support an operational mode, such as ICME. The center is looking at infrastructure to allow the models to move seamlessly into manufacturing. It is also working with Caltech/JPL on integrated systems design, which looks at how design models can inform the system and how to use models in production.

Dr. Davis said that smart manufacturing is based on a testbed approach. He explained that the SMLC’s portfolio of problems includes the following: smart machine operations, high-fidelity modeling, dynamic decisions, enterprise and supply chain decisions, and design and planning. All of these elements are data-intensive in some way, and an infrastructure is needed to manage these data appropriately.

Dr. Davis described smart manufacturing as a multilayered system, and he postulated that improvements can typically be made at points of handoff or seams. Seams can be between different departments or vendors, between designers and manufacturers, or between business systems and control/automation. At the lowest layer, the microlayer, the focus is on insertion, rapid qualification, ICME, and informing control systems. There is a short time constant associated with these functions. The next layer, the mesolayer, is a much larger space and consists of the operational decisions. The focus is on operational performance and goals, maintenance, dynamic trade-offs, and people. The upper layer, the macrolayer, focuses on supply chain information and transitions to outside the company. Dr. Davis pointed out that there are seams within each layer as well as seams across layers. The time constants are different across the different layers, which creates seams as well. It is important to orchestrate applications—that is, manage the work flow. This construct allows one to manage time (a “window of action”). The work flow can be analyzed and then used to generate projections about the output. It also allows for tracking and traceability.

Dr. Davis then defined smart manufacturing intelligence and work flow. Smart manufacturing intelligence is characterized by

- Applications that can share data, data that can share applications, and applications that can connect to applications to achieve horizontal enterprise views and actions.
- Orchestration of standardized decision work flows based on structured adaptation and autonomy.
- Actionable data, trust, and visibility across the supply chain.
- In-time, in-production qualification of materials, products, and actions.
• In-time, in-production, multidimensional (business, operations, supply chain, customer, maintenance, energy) performance and adaptation.
• Cross-company operational data to improve performance.
• Evolvable design models in manufacturing.

**FIGURE 3** Smart manufacturing work flow.

The sections circled in red at the bottom right are related to infrastructure.

SOURCE: Jim Davis, University of California, Los Angeles, presentation to the committee on February 5, 2014, Slide 15. Courtesy of Smart Manufacturing Leadership Coalition.

The SMLC also defined work flow, stating that the smart manufacturing work flow enables a dynamic orchestration of manufacturing steps across different time constants and seams, including the supply chain, without losing control of state. It is hosted on an interoperable, accessible, affordable, secure, and reliable hybrid cloud platform and supports commercial products and services, research and development needs, and academic interests.

Dr. Davis explained how smart manufacturing interfaces with the ISA-95 standard. He said that the SMLC takes advantage of standards but has not itself been involved in standards development. With respect to the definitions of ISA-95 Levels 0, 1, 2, or 4, the coalition is most interested in ISA-95 Level 3, which relates to manufacturing operations management. Level 3 seems to be a growing space with the most interfaces and thus offers the greatest opportunity for improvements in efficiency.

The SMLC looked at the main reasons the coalition stays together: Most SMLC members are interested in issues that extend beyond any one company. These issues are for example related to risk and organizational constraints. Data are part of each element.

The SMLC used a market-driven approach to identify its work flow, focusing on work flow as a service. The SMLC used the smartphone model of apps, toolkits, and so on. “App” is used here very broadly as a general-purpose term that allows systems to interface. The app model allows for enhanced flexibility. A work flow schematic is shown in Figure 3.
Dr. Davis explained that the SMLC uses a market-driven approach. Mapping and interface apps retrieve and manipulate data and map context. The apps should be put somewhere visible, and users should be able to understand how the apps are to be used and how well they perform. The SMLC is also interested in infrastructure that allows for the contribution of apps, broadening the space of innovation. The market is also driving standards-setting. Dr. Davis said this is equivalent to assembling a stack where the lower levels focus on the IoT. The SMLC focuses more on the middle layers, related to smart systems. Big data is the focus of the top of the stack; the amount of data increases as one rises through the stack. Currently, no one wants to share data, but there is a willingness to share apps and their use for different data applications. Dr. Davis pointed out that data valuation is a critical issue; when benefit is derived from the data, understanding the intellectual property becomes important, as does developing a business model to capitalize on that benefit.

Dr. Davis explained that there are many parallels in the health care system. The health care system also has value associated with patient data, as well as derivative intellectual property. One can think of companies as patients.

Dr. Davis noted that the app terminology relates smart manufacturing to smartphones. The chip layer in a smartphone is analogous to the SMLC work-flow-as-a-service layer; carriers match up to large manufacturers, because both groups address issues related to matching platforms and compatibility. Also, smartphones have free core apps as well as paid apps, just as in the smart manufacturing realm. One distinction is that the smart manufacturing community is linking the data flow between existing apps. The apps themselves are not new, but the toolkits (work flows assembled to have a certain function) are the most valuable to the manufacturing community.

Dr. Davis concluded by saying that smart manufacturing cannot be addressed piecemeal. The coalition remains cohesive in its attempt to take a comprehensive view of how to proceed in smart manufacturing, focusing on the architecture to enable and orchestrate the apps while allowing the marketplace to decide the use.

In the discussion period, a participant pointed out that smart manufacturing is customer-driven. In the defense community, the notion of a customer may be less clear. The military is a consumer, but not necessarily a customer. Dr. Davis responded that whereas the focus would have to be less centered on the consumer when applying this to the DOD, the notion of orchestrating and coordinating is still important. Another participant suggested it may be better to consider the defense industrial base as the customer.

Dr. McGrath made some specific points related to Kepler. He pointed out that the smart manufacturing work flow is based on Kepler, which is a scientific platform that has been used for many years for scientific work flows. He pointed out that Kepler currently seemed to be working well, but as the system becomes vendor-agnostic and moves into the open source, the specific platform may need to be reassessed. He pointed out that Kepler will need to be able to capture metadata.

Dr. Shafrik asked about the differences between small manufacturers and larger ones. Dr. Davis said that the smaller manufacturers do not have the time or money to address the issues related to smart manufacturing. If a smaller manufacturer is given the tools to manage information for multiple customers, it is likely to use them.

**DISCUSSION**

A participant asked about access to DOD data. The response was that AFOSR, ONR, DARPA, and others are exploring ways to share data generated under a DOD contract with the broader community.
Another participant asked about validating data, particularly when using advanced manufacturing techniques. Dr. Irving said that there is more emphasis right now on improving advanced manufacturing techniques; once the technique is refined, one can consider which process parameters to capture.

A participant argued that, for the small-scale researcher, data collection is outpacing computing capabilities. In the case of neutron beams, the bottleneck is processing the data rather than gaining access to beam lines. Continued progress is needed in to provide new data analysis programs.

A participant pointed out that an industrial manufacturer such as General Mills has commercial supply chains. However, the DOD does not. It can be difficult to keep all points of the DOD supply chain engaged when demand falls off. Another participant noted that an older NRC report commented on the dual-use industrial base (National Research Council, 1999). That report pointed out that, in a downturn, a company that is optimized for the defense industry cannot be readily commercial successful at the same time. There are no technical constraints, but the business models do not match well.

A participant concluded the discussion period by noting that the materials community has needs in all four V’s of big data (volume, velocity, variety, and veracity) but at different levels of urgency. The materials community may not be as limited by computational technology as other fields are, though there are still unresolved issues in how to manage and analyze data.

**Session 3: Big Data Issues in Manufacturing**

Session 3 of the workshop focused on the application of big data concepts in manufacturing. Adele Ratcliff, Office of the Secretary of Defense, had originally been scheduled to present during Session 3 but was unable to participate. Presentations were made in Session 3 by Jesse Margiotta, DARPA, and Wayne Ziegler, Army Research Laboratory.

**DATA NEEDS TO SUPPORT ICME DEVELOPMENT IN DARPA OPEN MANUFACTURING**

Jesse Margiotta, Technical Advisor, DARPA

Dr. Margiotta began by saying that today’s qualification and certification paradigm for parts and processes is fraught with difficulties. The methods are empirical, sequential, and iterative, leading to potentially prohibitive increases in cost and time. He said that the greatest challenge in qualification and certification, however, is its uncertainty. Dr. Margiotta pointed out that uncertainty in the qualification and certification process adds risk to a project, preventing new technologies from being incorporated into larger systems. As a result, the current qualification and certification paradigm creates a barrier to technology innovation and adoption. To counteract this, DARPA has begun the Open Manufacturing initiative; its main goal is to build and demonstrate a rapid qualification framework that aims to comprehensively capture, analyze, and control manufacturing variability. Dr Margiotta explained the guiding principles for DARPA’s Open Manufacturing program:

- Identify critical parameters, variation, and limits early in the process.
- Reduce testing and development iterations.
- Predict location-specific probabilistic performance.
- Build confidence in new technologies or qualification processes.
- Accelerate process maturity and systematic process reassessment.
Dr. Margiotta then described a project DARPA is developing with Honeywell Aerospace and several other team members. The project aims to develop rapid qualification of powder bed fusion additive manufacturing processes—in particular, direct metal laser sintering (DMLS). The general approach consists of the following elements: parameterize the manufacturing process; implement new sensors into the manufacturing process; incorporate an ICME construct that links process to materials to properties; and apply rigorous model verification and validation to understand the confidence limits. In this way, process parameters are linked to quantified, location-specific properties of the as-manufactured part. Dr. Margiotta showed a schematic of the critical elements for rapid qualification (Figure 4). The constituents to enable rapid qualification are shown in blue at the top of the figure. The supporting elements are shown below that; many of them—such as sensing, linking sensing capability to quality assurance, and microstructure property models—still need to be developed. Dr. Margiotta pointed out that the business cases and implementation plan are particularly important, as they will affect the usage and acceptance by the broader community. He stated that the architecture consists of increasing layers of complexity, including difficulties with the interfaces between different elements of the system.

**FIGURE 4 Critical elements of a rapid qualification system.**


Dr. Margiotta then explained the informatics associated with the additive manufacturing process. First, experiments are conducted to define the processing window, which is then refined through additional experiments to determine the optimal site within that window. This leads to a semioptimized process and overall improved material properties. The energy input density can also be measured and correlated with the quality of the consolidated material. In addition, the build chamber is instrumented to provide real-time monitoring of process parameters. The sensors have been able to capture a large quantity of high-fidelity data; at this point, about 1 TB of sensor data are collected for each DMLS build.

Dr. Margiotta then moved to the ICME construct, which uses process–microstructure–performance models to simulate the manufacturing process. Dr. Margiotta explained that the current simulation takes several days to a week to complete, which is much too long. These tools need to be further developed and simplified. The ICME construct consists of the following elements:
• Computationally intensive, physics-based models to simulate the manufacturing process. These models simulate the laser interaction with the powder bed, including thermal profiles and heating rates.
• Microstructural models to predict stresses, grain size, strain hardening, and other variables.
• Yield strength prediction tool.
• Uncertainty quantification to understand the relationship between processing and properties and the sensitivity of this relationship.

Dr. Margiotta said that the framework is the most critical element of the system. Once the general framework is in place, tools can be swapped in as they are developed. He noted that the tools are still under development and that much work remains.

Dr. Margiotta explained that the Open Manufacturing project was one of the first to extend verification and validation and uncertainty quantification to ICME processes. The Open Manufacturing project intends to draw on the work and standard practices developed in other fields in which simulation is well-developed and rigorously validated. These same methods can be transitioned into the materials and manufacturing arena.

A participant asked what is meant by "rapid" in this context, because the term can mean different things do different people. Dr. Margiotta responded that a typical qualification effort takes at least several years, with a long development effort prior to that before the qualification is begun. He was hesitant to associate a specific number with "rapid" but pointed out that there are significant time savings to be had in qualification efforts.

A participant also asked about the meaning of probabilistic design. Dr. Margiotta explained that one identifies the worst defects possible and puts them into the most critical locations, then, using the minimum material properties, designs the part so that it will be able to withstand that worst-case scenario. Alternatively, probability should be used to understand the likelihood of the defect, its location, and its effect and should then optimize the design accordingly.

Dr. McGrath asked about the data that have resulted from the Honeywell additive manufacturing project. Dr. Margiotta responded that the intent is for these programs to provide the data in a data archival tool, as is described in Mr. Ziegler’s talk, below). Access will be provided to other government agencies, with the details on broader distribution still to be determined. The process and materials data will not become proprietary.

Dr. Margiotta was asked if this constituted a "big data" problem. He responded that the project is generating considerable amounts of data, but it is not considered big data based on today’s definition. However, the materials manufacturing community does not currently have the ability to manage and analyze even this relatively modest amount of data.

Dr. Margiotta also pointed out that DARPA, along with ARL and other program partners, is developing methods to standardize data fields and metadata fields for materials and materials processing.

THE MATERIALS INFORMATION SYSTEM

Wayne Ziegler, Materials Engineer, Army Research Laboratory

Mr. Ziegler began his presentation by explaining the value of a material information system. He noted that, while it is not the case for this audience, he often needs to convince his listeners of the value of the materials information system
approach; people often are more concerned with intellectual property or do not understand the problems that currently exist. He pointed out that a materials information system does the following:

- Enables researchers to work more quickly and intelligently.
- Reduces duplication of effort in test and evaluation, which correspondingly reduces costs.
- Stops data loss and ensures data are available for the next generation.
- Improves data consistency and quality.
- Improves work processes and throughput.
- Accelerates implementation.

**FIGURE 5 Materials information system.**

**SOURCE:** Wayne Ziegler, Army Research Laboratory, presentation to the committee on February 6, 2014, Slide 4.

Mr. Ziegler said that a successful data management plan uses a systems engineering approach and includes four main components: capture, analyze, deploy, and maintain. He suggested that the DOD had historically had difficulty with maintaining programs that manage material and process data due to the challenges of a mobile workforce and shifting budget considerations. NASA also has a long history with strict materials data management, as several major catastrophes at NASA were related to materials issues. As a result, ARL is working with NASA to identify lessons learned and leverage NASA’s experience and IT infrastructure resources. Current challenges associated with the development of materials data management plans, both in industry and DOD, include these:

- Lack of direction.
- Lack of adequate resources.
- Lack of a return-on-investment business case.
- Lack of agreement: Not all companies or agencies believe that all data should be shared, and the cultural mindset needs to be changed.

Mr. Ziegler said that the goal is to build a DOD resource for materials and process information. The DOD resource, Materials Selection and Analysis Tool (MSAT), is currently hosted by NASA as an independent component of the NASA
Materials and Processes Technical Information System (MAPTIS) system. He went on to say that the MSAT program has a strong partnership with DARPA and its Open Manufacturing program (see Dr. Margiotta’s presentation, above).

Figure 5 shows a vision of a materials information system. It begins with experimental methods; Mr. Ziegler pointed out that we tend to lose metadata in this area, and experimental methods and results should be part of an integrated database structure. Metadata include information related to testing conditions and program information necessary for data sets to be completely understood and if necessary validated through additional testing. Data mining techniques are then applied to the data and the mined data are used to inform models. This process is iterative, and requires the tracking of data pedigree, in other words, data about the data. Mr. Ziegler noted that as much as half of the data can be pedigree information. Relevant metadata are needed to compare data across data sets. Mr. Ziegler indicated that the Open Manufacturing project follows this general form.

Mr. Ziegler pointed out that MSAT is called a selection and analysis tool since its initial focus is making programmatic or research decisions based on the data sets available in a robust and timely way. MSAT has a wide approach to application, modeling, resource management, process approval, and improvement.

Mr. Ziegler then argued that there needs be a cultural shift in work flow management. The traditional work flow paradigm is to execute a task, collect and extract data, return a bigger data set, and pair it with separately recorded information about the process. However, when the collection of data is separated from the collection of process information, fidelity drops. Mr. Ziegler argued in favor of a new work flow paradigm, whereby when a process is executed, the data and the metadata are collected simultaneously. When data are extracted from this set, there would already be a link that coupled them in any future data processing. He went on to describe the steps in the materials information system:

1. Define data sets. Mr. Ziegler pointed out that any data that can be collected in a reasonable fashion should be collected, as it might prove useful later. ARL is still addressing this first step, and Mr. Ziegler explained that the data collection decisions are iterative; once they have started collecting a particular data set, researchers will likely determine that they will need other data as well.

2. Define data management schema. This process looks at how to organize and arrange the data. It is also an iterative process.

3. Develop import templates. Mr. Ziegler noted that ARL currently uses Excel; the templates are in a comma-separated values format so that they are software-agnostic. Several companies provide commercial data management packages or modeling software, and the objective is to build templates that can interface with a variety of commercial software.

4. Use templates to import data.

5. Manage data. This step includes defining access control and conducting verification and validation.

6. Define the use cases. The resulting information is used to define output templates.

Mr. Ziegler suggested that Steps 1, 5, and 6 are the most critical for the users. Once the use case is known, then the data can be exported in a useful way. Mr. Ziegler then discussed several technical considerations, including the following:
• Defining the main function of the system: Is it capturing a manufacturing process, or does it focus on identifying material properties?
• Integrating the system with existing systems and workplace practices with minimum impact.
• Understanding and how system users will use the data.
• Data flow through the system. Mr. Ziegler noted that in the materials science and engineering technology area this is not “big data” (yet), though it is on a large scale.
• The type of information to be handled.
• System setup, deployment, and maintenance.
• Responsibility for, and ownership of, the various system components. Mr. Ziegler said that this can be a contentious issue, as data have value. Defining data access may not always be technically challenging, but it can be a policy challenge.

Mr. Ziegler concluded by noting some practical considerations, including these:
• Not every user is an expert, so the user interface becomes critical.
• Materials and process data are usually incomplete.
• Data have value; access control is critical and potentially contentious.
• It takes time to rationalize and consolidate data. The better the system is at collection, the better it will be at consolidation.
• Data end users need data in diverse places and formats.
• Materials information systems need end users. Mr. Ziegler argued that there is no value in the system if it does not have end users. It can be challenging to identify and engage users.
• Designing a system from scratch is impractical.

A participant indicated that data ownership can be an obstacle to data sharing. He said that DOD contracts have many data requirements in them, and that aspect needs to be managed on the contractual side to ensure that the requirements are not cost prohibitive. He pointed out that there is a responsibility for sharing data among material suppliers, original equipment manufacturers, and the government. In some cases, a supplier provides a material but no corresponding metadata. Agreements with suppliers can take 1-2 years to develop, which slows down innovation. Mr. Ziegler agreed that acquisition is an important element, though outside the scope of ARL’s mission and activities.

Dr. McGrath asked for clarification on MSAT. Is it a tool for materials selection, with a correspondingly fairly limited user community? Or is it an element of a larger system within a larger community, with a framework surrounding it? What is the plan for scaling up past the Open Manufacturing project? Mr. Ziegler said that MSAT is both a materials selection tool and part of a larger system. MSAT’s current focus is on where to store materials and processes and how to develop a clear interface with the modeling community.

The discussion then turned to standards. A participant stressed that the process for developing standard terminology is very difficult and slow. There is an ASTM committee for standards in this area. Companies do not like to fund their employees to do this type of activity, however, and the ASTM committee terminated its efforts because of insufficient community funding. Also, companies are not interested in attaching themselves to a certain format, as they are concerned they will be forced to share data. They prefer to keep information proprietary in their own formats. A few participants noted that the culture among researchers and companies is such that materials data are considered a competitive edge, and companies want to protect their intellectual property.
DISCUSSION

Valerie Browning, from ValTech Solutions LLC, opened the discussion by noting that the workshop speakers thus far have discussed materials challenges in variety and veracity, but not volume or velocity. In materials, therefore, it may be more important to think about information rather than big data. The materials area has an extra layer of extraction or analytics that is unique to this community. The questions then become, Who is responsible for developing the analytics? How can we manage work flow on different time scales? and Who owns the analytics?

Another participant agreed that there is a data problem in materials science, but not a big data problem. He said that the data problem seems to center on data collection and the lack of sharing. He suggested that a mandate is necessary stating that any government-funded data must be put into a standard format, a step that is being considered by NSF and DOE. A DOD participant said that DOD has explored the idea of such a mandate but finds it time consuming and expensive, and, in the end, concludes that the cost may outweigh the benefit. He indicated that the issue is more than one of data format; it includes questions about who owns and maintains data and where the information should reside. Someone else remarked that the NSF repository is not user friendly. Another participant pointed out that it is fairly common for universities to have permanent storage facilities available and gave the Deep Blue program at the University of Michigan as an example. However, other participants argued that these programs are expensive and do not always include metadata.

One participant believed that there is a data collection problem in manufacturing. Manufacturers need to organize data definitions, contextuals, the meaning of operations, and connect with ICME. Dr. Davis brought up the analogy of health care, and finds many parallels in the health care shift to digital patient records. Another participant noted that advanced manufacturing can be done on a small scale that is not necessarily part of a large corporation. This could add a layer of complexity, as a small player may not be interested in negotiating business plans.

Session 4: The Way Ahead

The final session of the workshop focused on the federal government’s present and future projects and plans related to data in materials and manufacturing. Presentations were made in Session 4 by Julie Christodoulou, Office of Naval Research, and Michael Stebbins, Office of Science and Technology Policy.

LIGHTWEIGHT AND MODERN METALS MANUFACTURING INNOVATION INSTITUTE: IMPLICATIONS FOR MATERIALS, MANUFACTURING, AND DATA

Julie Christodoulou, Director, Naval Materials Division, Office of Naval Research

Dr. Christodoulou began her presentation by describing the Lightweight and Modern Metals Manufacturing Innovation Institute, known as the LM3I Institute (LM3II). The LM3II provides an opportunity to integrate materials into emerging manufacturing technologies in a meaningful way to ultimately produce better-performing systems and components for military and civilian applications, and to do so faster and more affordably. LM3II is part of a national network of institutes that focus on topics of national import and benefit. It will have many partners, and DOD is leading the effort. Dr. Christodoulou said that metals are uniquely dual-use. There is no conflict between military applications and commercial value; DOD welcomes commercial applications of materials technologies, because a broader marketplace will help reduce costs for the military. The goal of the LM3II is to demonstrate advanced manufacturing capabilities to enable systems that are lightweight, reliable, survivable, fuel efficient, affordable, and flexible while designing lightweight metals...
and applications that use them. Lightweighting has a near-term impact on fuel efficiency, flexibility, and overall costs for naval systems and platforms, and Dr. Christodoulou pointed out that many of these benefits were discussed in the National Research Council’s report on lightweighting (NRC, 2012). She said that LM3II focuses on aluminum, titanium, magnesium, and other modern metals and on manufacturing readiness levels (MRLs) 4-7. It is interested in incorporating ICME to shorten the design–production cycle. The goal is to have the program be self-sufficient in 5 years, with no dedicated funding after that time. Achieving this goal will require the identification of transition partners.

She then briefly described the ICME process (see Box 2 in Dr. Ward’s summary for more information about ICME). Dr. Christodoulou argued that ICME is an important driver in the acceleration of development times.

She then briefly discussed DARPA’s Accelerated Insertion of Materials (AIM) program. The AIM program was established to apply ICME to specific applications in the aircraft engine industry so as to reduce development time and costs. The AIM program focused on the designer’s toolbox, including knowledge, heuristics, and collections of data sets. Dr. Christodoulou stressed the importance of being able to use that set of knowledge and heuristics to exert greater control over a material’s structure and properties. She pointed out that ICME can help materials scientists communicate better with part designers, enabling them to take advantage of the complexity of a material’s microstructure to design a better component.

Note: MRL is a scale on which to assess manufacturing maturity and risk using a standard set of criteria. MRLs range from 0 (basic manufacturing implications identified) to 10 (full rate production demonstrated). See http://www.dodmrl.com for more information about MRL definitions. Accessed February 27, 2014.

Dr. Christodoulou then provided an example of a project with exceptionally complex data. The example related to grain growth in a titanium alloy. The project, led by Lauridsen and Voorhees, used high-fidelity three-dimensional synchrotron imaging techniques and phase field modeling to capture anisotropy effects. The project used successive images to study the progression of grain boundaries to understand the microstructure. While the data collection process took about a week, the data analysis took over 2 years because of the exceedingly complex data (Poulsen, Voorhees, and Lauridsen, 2013).

Dr. Christodoulou also stressed the importance of mixing models at multiple length scales to predict microstructural evolution and resultant localized properties. She pointed to the work of Li and Wang in phase field models of microstructure evolution. Phase field models developed at different length scales provide a useful means to capture fundamental insights into phase transformation and deformation mechanisms and to establish physics-based models for engineering applications. Again, in order to do this, large and realistic volume elements must be explored. Materials naturally have a large amount of heterogeneity associated with them, and that space needs to be examined. The data need to be very rich in detail in order to be able to create an accurate model. These systems also require powerful analysis tools to acquire the data and develop models that are correspondingly rich in detail.

Dr. Christodoulou then discussed the Advanced Manufacturing Partnership and the MGI. The two initiatives were announced jointly in June 2011 and are synergistic. (See Box 1, in Dr. Ward’s presentation, above, for more information about MGI.) Dr. Christodoulou pointed out that ICME is the primary tool to address the challenges presented in the MGI, and MGI is therefore the entity to enable widespread adoption of ICME—to make best practice become common practice. She also reminded the audience that there is a continuum in materials development: The phases of materials development (discovery, development, property optimization, systems design and integration, certification,
manufacturing, and deployment) are all synergistic, and researchers in each phase need to be able to interact using a common language.

Dr. Christodoulou then discussed the National Network for Manufacturing Innovation (NNMI), a program that is designed to look at the interfaces among industry, academia, and government partners to identify gaps and provide the missing physical and intellectual middle ground. NNMI is a technology incubator for small organizations or multiagency groups. A schematic of the vision for NNMI is shown in Figure 6. Dr. Christodoulou described the NNMI as an incubator space, to which smaller or multicomponent groups can bring a problem or idea and test its viability. She noted that the NNMI seeks to provide opportunities for education at all levels, including sabbaticals, student co-ops, and internships.

Returning to the goals of the LM3II, Dr. Christodoulou emphasized the importance of bringing in industrial counterparts early in the design and development process; she noted that this is one of the central requirements for ICME. The transition pathways will then be established from the very onset of a project. She reminded the audience that LM3II is looking for self-sustainment—that is, independence from dedicated federal funding—in 5 years, so transition opportunities must be explored now. She believes that the LM3II will be formalized in the near future.

Dr. McGrath asked about the LM3II’s expectations and whether it anticipates becoming like the open platform of Open Manufacturing. Dr. Christodoulou responded that LM3II would be a complex, public-private partnership for the next 5 years, with $70 million total investment from the government over that time and with expectations of more than that in matching funds. The Institute is currently sorting through issues related to levels of engagement and intellectual property. She said that each project would likely have its own unique intellectual property arrangements, although the approach to developing those arrangements can be established early. The LM3II is working toward common data formats, while making sure to understand the different architectures under development.

**FIGURE 6 Elements of the NNMI**

SOURCE: Julie Christodoulou, Office of Naval Research, presentation to the committee on February 6, 2014, Slide 28.

Dr. Schafrik asked about the GE example, which consisted of a project in solid-state welding in titanium. He noted that there are unusual phases at the interface zone. He asked if those were under investigation in this project, or whether the project was considering only the monolithic material. Dr. Christodoulou responded that they hope to look broadly at manufacturing technologies, including joining.
A participant asked if Dr. Christodoulou has considered cartography to represent microstructure. Dr. Christodoulou said that some element of mapping has to be undertaken to represent microstructure. There is a need for standards to capture the properties. Another participant pointed out that the materials and manufacturing community borrows from the medical community in imaging. Dr. Christodoulou agreed and noted that the digital data set is not simply a digital image. Material properties, performance, and structure evolution must be quantified, along with any changes to those variables. Still another participant said that collections of microstructures tend to focus on small-scale changes but that it is important to simultaneously capture changes on longer scales.

**DIRECTION OF POLICY**

Michael Stebbins, Assistant Director for Biotechnology, Science Division, White House Office of Science and Technology Policy

Dr. Stebbins explained that he is the White House lead on openness and access to scientific data. The White House activity encompasses all fields of scientific study, and it has direct relevance to the materials community. He pointed out that many workshop participants noted that material science does not have what is generally considered a “big data” problem.” However, he would like for there to be one: He would prefer to see that the materials science community has too much data rather than not enough. Dr. Stebbins pointed out that, looking broadly across all scientific disciplines, it is difficult, if not impossible, to access the data that underpin figures in scientific journals. There were historical reasons for this, he argued, not practical limitations, as data storage is now inexpensive and widely available. The obstacles that prevent data access include cultural mindset, lack of suitable storage locations, lack of data standards, and insufficient funding for managing data sets. Dr. Stebbins pointed out that the biomedical research community has been approaching a crisis in that it has become virtually impossible to reproduce biomedical results. This is not because of widespread fraud, but because of a lack of access to protocols, underlying data, and process information. The biomedical community has developed good monotypic databases, with analytic tools superimposed on them. As a result, the community is able to share insights and accelerate development. Dr. Stebbins argued that genomics is a good example of a field in which data sharing is particularly critical. He added that, while monotypic databases are important and should continue to be used, something more may be necessary for a large corpus of data and said that a 3-4 PB per year database is a good prototype starting size. The White House is in the process of working with the scientific journals to require that any article published in them have referenced data with a persistent identifier, including metadata and proper credit.

Dr. Stebbins pointed out that the only “currency” available today to a research scientist is the scientific publication in a journal. It may be possible to add a new currency: the actual data. He said that frequently a single paper may have many coauthors, and the individual contribution of each author is unclear. This is particularly true in certain fields, such as high-energy physics. Published data may be able to clarify the individual’s role. In some fields, publishing the data may not be feasible or useful; in others, such as biology, it will dramatically accelerate development and allow for reproducibility. He asked the workshop participants if materials science would find the publication of data useful. Any potential advances, unintended consequences, or possible hazards in materials science should be explored now, before engaging the relevant publishers.

Dr. Stebbins pointed out that the government has had relatively little impact on increasing data sharing. By contrast, federal laws tend to focus on preventing the government from forcing the sharing of results. The Bayh-Dole Act, 20 for instance, encourages researchers to retain their discoveries for the purpose of commercialization. However, the
The government has not had much experience in the promotion of data sharing. The most successful instances of data sharing have been driven by publishers, who seek to ensure data quality in their publications.

A participant pointed out that the National Nanotechnology Initiative (NNI) has a focus area in informatics and is working with relevant journals on data access as well. Dr. Stebbins said that it would be important for conversations to be coordinated with White House policy and with the bigger picture surrounding information sharing.

Another participant pointed out that universities are struggling with open publishing vs. licensed publishing. Universities must pay to access the journals. These costs may be amplified if the access also suddenly includes additional costs for data access as well. Dr. Stebbins said that the initial pilot programs OSTP is considering will be free. The researchers (not the publishers) will deposit their data with a registered third party. In time, standards will develop for the database, and publishers may require a specific data management solution that costs additional money. As an example, the journal Nature has an agreement with the company FigShare. Participants worried that this additional cost for having FigShare manage the data is likely to be passed on to the authors, and publishing in Nature is already quite expensive. A high-end journal such as Nature is likely to succeed with this model, but other journals may find that additional data costs will discourage publication submissions. Dr. Stebbins pointed out that other private sector entities will likely become more engaged in data storage and management, including the nonprofit community. A participant asked who would select the data storage site. Dr. Stebbins replied that the journal could say that it has a preferred database or a partnership with a particular data-sharing site. Alternatively, the journal could establish the data standard. Data longevity will need to part of the standard; longevity would need to be guaranteed. Self-storage is unlikely to meet any longevity standard.

Various participants discussed questions about access to data that still needed to be addressed, which data set(s) needed to be provided, and data standards. Someone argued that it would not be sufficient to upload a data set; authors must upload models as well. Dr. Stebbins agreed that models should be part of the metadata, including specific algorithms and software versions. Someone else pointed out that publishers are a good point of leverage in the academic community. In communities such as pharmaceuticals and other industries, however, publication is not a high priority. The data in those instances are used to build business and are considered proprietary. It becomes difficult, if not impossible, for the outside community to gauge the quality of those data. The participant supposed that it would be very challenging to try to move into the commercial space with this proposal.

Wayne Ziegler and Chuck Ward both said that this open data initiative is consistent with and complementary to the activities in ARL and AFRL. Dr. Ward said that both their initiatives have been in communication with NIST and the relevant journals. He said that they are working to develop standards for CALPHAD data. They will next work with issues related to more complex data. He noted that it is fairly easy to come to a set of data standards for crystallography and CALPHAD, but other more complex data sets, such as three-dimensional data, will be more difficult.

A participant voiced concern that materials research has a lot of associated information or metadata, including sample preparation, and tracking the amount of metadata associated with an individual sample may be onerous. Dr. Stebbins pointed out that while this could be true it was still his belief that it is tractable.

Footnotes
For a full description of the four V’s, see the section “IBM and Big Data” in Session 1.


OODT can be explored online at http://oodt.apache.org/, accessed February 20, 2014.


Watson is an artificially intelligent computer system capable of answering questions posed in natural language. At one point Watson had access to 200 million pages of structured and unstructured content in a database consuming 4 TB of disk storage.


Hadoop, developed by the Apache Software Foundation, is an open-source software framework for the storage and processing of large-scale data sets on clusters of distributed computers. See http://hadoop.apache.org/ for more information. Accessed February 20, 2014.

The Dryad Digital Repository is an open source to make available the data that underlie scientific publications. See http://datadryad.org for more information. Accessed February 24, 2014.


BMI Corporation is an engineering services firm based in Greenville, South Carolina.


CALPHAD refers to the CALculation of PHAse diagrams, a computational method of modeling the thermodynamic properties of materials.


IoT is the network of uniquely identifiable physical objects (such as sensors or actuators) embedded throughout a network structure.

ISA-95 is an international standard for the integration of control systems for manufacturing and processing operations. See http://www.isa-95.com for more information, accessed February 25, 2014.

Kepler is a free, open source, scientific work flow tool. See http://kepler-project.org for more information, accessed February 25, 2014.

See http://deepblue.lib.umich.edu/ for more information, accessed February 26, 2014.

MRL is a scale on which to assess manufacturing maturity and risk using a standard set of criteria. MRLs range from 0 (basic manufacturing implications identified) to 10 (full rate production demonstrated). See http://www.dodmrli.com for more information about MRL definitions. Accessed February 27, 2014.
The Bayh-Dole Act of 1980 amended patent and trademark law to permit universities and small businesses to retain intellectual property rights to work conducted using federal funding. (Public Law 96-517).

References


A Workshop Statement of Task

An ad hoc committee will convene a series of three 2-day public workshops to discuss issues in defense materials, manufacturing and infrastructure including: 1) Globalization of Defense Materials and Manufacturing; 2) Big Data in Materials Research and Development; and 3) Materials State Awareness. The committee will develop the agendas for the workshops, select and invite speakers and discussants, and moderate the discussions. The workshops will use a mix of individual presentations, panels, and question-and-answer sessions to develop an understanding of the relevant issues. The workshop topics will highlight some recent developments in the fields. Key stakeholders will be identified and invited to participate. Individually authored workshop summaries will be prepared separately by a designated rapporteur after each workshop in this series.

B Workshop Participants

Bharat Bhushan, Ohio State University
Valerie Browning, ValTech Solutions, LLC
Dennis Chamot, National Research Council (retired)
Julie Christodoulou, Office of Naval Research
Dan Crichton, NASA Jet Propulsion Laboratory
Jim Davis, University of California at Los Angeles
Jesus de la Garza, Virginia Tech
Robert Dowding, U.S. Army Research Laboratory
Barry Farmer, U.S. Air Force Research Laboratory
David Forrest, Department of Energy
Steven Freiman, Consultant
Lisa Friedersdorf, National Nanotechnology Coordination Office
Rosario Gerhardt, Georgia Institute of Technology
Natalie Gluck, Institute for Defense Analyses
Ken Hix, GE Aviation
Russell Irving, GE
Robert Latiff, R. Latiff Associates
Alexis Lewis, Naval Research Laboratory
Steven Linder, Office of the Secretary of Defense
Kenny Lipkowitz, Office of Naval Research
Jesse Margiotta, Defense Advanced Research Projects Agency
Thom Mason, Oak Ridge National Laboratory
Suveen Mathaudhu, North Carolina State University
Peter Matic, Naval Research Laboratory
Anne Matsuura, Optical Society of America
Michael McGrath, Analytic Services, Inc.
Heather Meeks, Defense Threat Reduction Agency
Jed Pitera, IBM
Ward Plummer, Louisiana State University
Siddiq Qidwai, Naval Research Laboratory
Adele Ratcliff, Department of Defense, OSD, Manufacturing Technology
Mike Rigdon, Consultant
John Rumble, R&R Data Services
Robert Schafrik, GE Aviation
Dave Shepherd, Department of Homeland Security
Lewis Sloter, Office of the Assistant Secretary of Defense for Research and Engineering
Michael Stebbins, White House Office of Science and Technology Policy
David Stepp, U.S. Army Research Laboratory
Denise Swink, Private Consultant
Galip Ulsoy, University of Michigan
Hayden Wadley, University of Virginia
Chuck Ward, U.S. Air Force Research Laboratory
Scott Weidman, National Research Council
Jennifer Wolk, Naval Surface Warfare Center
Jeff Zabinski, U.S. Army Research Laboratory
Wayne Ziegler, U.S. Army Research Laboratory

C Workshop Agenda
FEBRUARY 5, 2014

8:30 am
Welcome and Meeting Objectives
Michael McGrath, Chair and Robert Schafrik, Vice Chair

9:00 Introduction to Big Data

Frontiers in Massive Data Analysis and Their Implementation: Daniel Crichton, JPL

9:40 IBM and Big Data: Jed Pitera, IBM

10:40 Bio and Big Data: Dave Shepherd, DHS

11:20 Discussion

1:00 pm Big Data Issues in Materials R&D

1:40 Physics in Big Data: Thom Mason, ORNL

2:20 MGI and Big Data Formats: Chuck Ward, Air Force

3:20 GE’s Implementation of ICME for Materials Data: Rusty Irving, GE

4:00 The Smart Manufacturing Leadership Coalition and Big Data: Jim Davis, UCLA

Discussion

FEBRUARY 6, 2014

8:30 am Welcome and What We Heard Yesterday: Mike McGrath, Chair

8:40 Manufacturing Issues in Big Data

9:20 Open Manufacturing/Honeywell Data Analysis: Jesse Margiotta, DARPA

10:20 MSAT/MAPTIS Data Archiving and Process and Materials Data: Wayne Ziegler, ARL

Discussion

11:00 The Way Ahead

11:30 The Lightweight Materials Institute: Julie Christodoulou, ONR

1:00 pm The Director of Policy: Michael Stebbins, OSTP

2:10 Discussion
3:00 Closed Session—Planning

Workshop Adjourns

### D Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADARA</td>
<td>Accelerating Data Acquisition, Reduction, and Analysis</td>
</tr>
<tr>
<td>AFOSR</td>
<td>Air Force Office of Scientific Research</td>
</tr>
<tr>
<td>AIM</td>
<td>Accelerated Insertion of Materials</td>
</tr>
<tr>
<td>CALPHAD</td>
<td>Calculation of Phase Diagrams</td>
</tr>
<tr>
<td>DARPA</td>
<td>Defense Advanced Research Projects Agency</td>
</tr>
<tr>
<td>DMMI</td>
<td>Defense Materials, Manufacturing, and Infrastructure</td>
</tr>
<tr>
<td>DOD</td>
<td>U.S. Department of Defense</td>
</tr>
<tr>
<td>DOE</td>
<td>U.S. Department of Energy</td>
</tr>
<tr>
<td>FDA</td>
<td>Food and Drug Administration</td>
</tr>
<tr>
<td>GE</td>
<td>General Electric</td>
</tr>
<tr>
<td>HIV</td>
<td>human immunodeficiency virus</td>
</tr>
<tr>
<td>ICME</td>
<td>Integrated Computational Materials for Engineering</td>
</tr>
<tr>
<td>IoT</td>
<td>Internet of Things</td>
</tr>
<tr>
<td>IPCC</td>
<td>Intergovernmental Panel on Climate Change</td>
</tr>
<tr>
<td>JPL</td>
<td>Jet Propulsion Laboratory</td>
</tr>
<tr>
<td>LM3II</td>
<td>Lightweight and Modern Metals Manufacturing Innovation Institute</td>
</tr>
<tr>
<td>MAPTIS</td>
<td>Materials and Processes Technical Information System</td>
</tr>
<tr>
<td>MDF</td>
<td>Manufacturing Demonstration Facility</td>
</tr>
<tr>
<td>MGI</td>
<td>Materials Genome Initiative</td>
</tr>
<tr>
<td>MSAT</td>
<td>Materials Selection and Analysis Tool</td>
</tr>
<tr>
<td>NASA</td>
<td>National Aeronautics and Space Administration</td>
</tr>
<tr>
<td>NIH</td>
<td>National Institutes of Health</td>
</tr>
<tr>
<td>NIST</td>
<td>National Institute for Standards and Technology</td>
</tr>
<tr>
<td>NNI</td>
<td>National Nanotechnology Initiative</td>
</tr>
<tr>
<td>NNMI</td>
<td>National Network for Manufacturing Innovation</td>
</tr>
<tr>
<td>NSF</td>
<td>National Science Foundation</td>
</tr>
<tr>
<td>OODT</td>
<td>Object Oriented Data Technology</td>
</tr>
<tr>
<td>ONR</td>
<td>Office of Naval Research</td>
</tr>
<tr>
<td>ORNL</td>
<td>Oak Ridge National Laboratory</td>
</tr>
<tr>
<td>OSTP</td>
<td>Office of Science and Technology Policy (White House)</td>
</tr>
<tr>
<td>SARS</td>
<td>severe acute respiratory syndrome</td>
</tr>
<tr>
<td>SMLC</td>
<td>Smart Manufacturing Leadership Coalition</td>
</tr>
</tbody>
</table>

**NEXT**

[https://semanticommunity.info/Data_Science/Data_Science_for_the_National_Big_Data_R_and_D_Initiative](https://semanticommunity.info/Data_Science/Data_Science_for_the_National_Big_Data_R_and_D_Initiative)

Updated: Sun, 14 Jul 2019 10:27:14 GMT

Powered by [Quri](https://www.quri.com)