Making the Most of Big Data

Story
Ontology and Big Data
FASTER CoP
Other
DoD Architecture Framework (DoDAF) Version 2.0 Plenary
Architecture Information Exchange Tools: Increasing the Rate of Change
Seeing Through Cloud Computing
History of Federally Funded Research and Development Centers (FFRDCs)
EarthCube - National Data Infrastructure for Earth System Science
Abstract
About The Speaker
Data on Data
Abstract
About The Speaker
Presenting PAGES
Abstract
About The Speaker
Cloud Computing and People with Disabilities
Abstract
About The Speaker
Cloud & Accessibility Workshop
Workshop Information
ITU-T X.1255
Abstract
About The Speaker
Cloud Service Level Agreements
Abstract
About The Speaker
Opening Government Data and Creating Public Value

Abstract

About The Speaker

Update on CASC HPC & Cyberinfrastructure Developments

Abstract

About The Speaker

Constructing High-Quality Cloud Service Level Agreements (SLAs)

Workshop Information

Data Infrastructure Building Blocks (DIBBs)

Abstract

About The Speaker

NEXT

Story

JHU Genetics

Linked TCGA: A clinically relevant 30 billion triple Dataset

Data to Knowledge to Action Event Highlights Innovative Collaborations to Benefit Americans

Building virtual clinical trials

Improving access to clinical trials

Bringing NASA data down to Earth

Applying data analytics to society’s problem-solvers

Full list of commitments presented at today’s event

Fact sheet on Data to Knowledge to Action: Progress by Federal Agencies

More information about NITRD

NITRD Supplement to the President’s Budget

Strategic Priorities

Current and Planned Coordination Activities

Data to Knowledge to Action Event Takes Big Data Initiatives to Innovative Heights

Building New Partnerships

100 Welcome

115 Building New Partnerships

200 A Bold New Collaboration

220 Coffee served in the Pavilion Foyer

230 Partnership Panel Discussion

330 Coffee served in the Meridian Foyer

345 Breakout Sessions

Education and Workforce Development
Government role 2
March 29, 2012 announcement
All Hands on Deck
Examples of actions to advance Big Data 1
Examples of actions to advance Big Data 2
Examples of actions to advance Big Data 3

The Promise of Big Data

Big Data Partners Workshop remarks by Suzi Iacono

Title Slide
Big Data Workshop 1
Big Data Workshop 2
Big Data Workshop 3
Multi-stakeholder Partnerships

Charge
Avoid these Idea Killers
Logistics
Parting Words

Panel Speaker Presentations

Building a Data-Driven Society
Data Science & Multi-Stakeholder Partnerships
Mark Peterson - Team Lead, Data and Analytics, Office of Science & Technology U.S. Agency for International Development
RENCI Collaborations and Consortia to Advance Data Science

Lightning Round Presentations

Genomes Galore – Big Data Analytics for High Throughput DNA Sequencing
Write Optimization: The Future of Big Data Storage
MIT Big Data Initiative at CSAIL
Big Data Science for the Masses: start small. Think Big
Self-Generated Health Information Exchange
Open Science for Big Data
Dataverse for Big Data
University of Kentucky
Synergistic Co-Design for BIG DATA
Data Science Incubator
MITRE’s Big Data Analytics Activities
Globus Online
Center for Urban Science and Progress
Confidentiality and Data Access in the Use of Big Data: Theory and Practical Approaches
Robert Wood Johnson Foundation
NELL: Learning to Read the Web
Microsoft Projects of Interest
SGI Solutions for Big Data
Data Bridge: Solving the First & Last Mile Problems in Big Data
IBM Brings Big Data Skills to the Classroom and Workforce
Analyzing (Big) Data Boxes: Multi-way CS of Tensors
Myria: Foundations and Systems for Big Data Management
Projects of the NRC Board on Research Data & Info, and US CODATA
Analytical Approaches to Massive Data Analysis
Integrating Humans, Machines and Networks: A Global Review of Data-to-Decision Technologies
Big Data Plus Cyber Security
The Center for Dynamic Data Analytics
National Weather Service for Infectious Disease
DOE Leadership Computing supports Big Data

Data Sharing and Metadata Curation: Obstacles and Strategies

Introduction
Executive Summary

Discussion and Presentation Summary
Practitioners' Perspectives
Trans-disciplinary Community Perspectives:
Open Discussion: Barriers and Opportunities:

Summary and Wrap-up:
  Workshop Summary and Wrap-Up
  With Big Data comes big risk
  Aligning metadata across pre-aligned domains?
  The sociological problem is as hard as than the technological problem
  Areas of Agreement?
  So what next?

Participant List
Organizing Committee
Glossary of Acronyms
Story

Data Science Mining of the FASTER CoP

Background: Followed the FASTER CoP activities (see summary below), written stories at the request of Dr. George Strawn (e.g. Theresa Pardo on Open Government Data), and invited the FASTER CoP leaders to do a Joint Meetup on September 8th (see proposed agenda).

Purpose: To do data science mining to identify big data projects for our Federal Big Data Working Group Meetup Data Science Teams to work on.

Example: Dr. Alexander Szalay who shared his perspectives regarding the Data Infrastructure Building Blocks (DIBBs) program on August 19th. I mentioned the astronomy use case pilot we are doing with Professor Kirk Borne involving ontology, graph computing, and SciDB.org and he thought it was a great idea and would like to follow up with us. So I think I will do a joint Meetup with Dr. Alexander Szalay (LINK TO NIST DATA SCIENCE WORKSHOP SLIDES) and Professor Kirk Borne to discuss our initial pilot results.

Specifics: Dr. Alexander Szalay said he liked our explanations of data science data publications because they would largely solve the metadata problem and would look at Spotfire (Dr. Ben Shneiderman’s invention) as an example of a tool that is "both a microscope and a telescope for big data".

The Big Data Senior Steering Group is currently working on developing a Big Data Strategic Plan for the Federal Government and have not yet gotten to the point where they are seeking external expertise on particular aspects; workforce development being one of them. MY NOTE: ADD WHAT HAS BEEN SAID PUBLICALLY BY THE WHITE HOUSE - I HAVE ATTENDED AND WRITTEN ABOUT THE BIG DATA AND PRIVACY WORKSHOPS AND REPORTS RECENTLY

NOTES ON EMAILS

SUMMARY TABLE ON FASTER CoP: I THINK WE COVER MOST OF THESE ALREADY: Cloud Computing, EarthCube, Research Objects, FFRDCs, etc.

MORE TO FOLLOW

Ontology and Big Data

Soliciting comments by proposing a framework for the issues of Track 3 Challenge
- Ernest Lucier, Coordinator, SDP and FASTER (Contractor), National Coordination Office for Networking and Information Technology Research and Development, 4201 Wilson Boulevard, Suite II-405, Arlington, VA 22230

January 24, 2012

Proposed framework for Track 3 Challenge - Ontology and Big Data
People in the domains of science, software engineering, computer science, etc. can benefit from a combined knowledge of their domain and application of ontology-based technologies. A combined understanding of these domains and ontology-based technologies may encourage the growth of technology.

Problems

- Programmers are not able to optimize the use of unstructured data for scientists and engineers
- Scientists without ontology training use brute force programming – can be inefficient and scientists and engineers are not aware of options and capabilities using ontology-based technologies
- Science and ontology-based technology evolution is slow or non-existent
- Is technological growth constrained by the shortage of qualified ontologists?

Goals

- Enable scientist to make maximum use of big data
- Enable scientist to understand the potential of ontology-based systems integration
- Enable ontologists to understand scientist needs

Needs

- The skills most needed today include a combined understanding of a scientific or engineering discipline and knowledge of ontology-based technologies.

Output (proposed)

- NITRD Big Data Challenge
- Other challenges

Resources

Links

- NASA-Harvard Center for Excellence in Collaborative Innovation
  - [http://www.nasa.gov/offices/COECl/index.html](http://www.nasa.gov/offices/COECl/index.html)
- Prizes & Challenges Community of Practice activities

Documents

- Future of Software Engineering Research (FoSER) report

Other

DoD Architecture Framework (DoDAF) Version 2.0 Plenary
January 27, 2011, at MITRE in McLean, Virginia

Architecture Information Exchange Tools: Increasing the Rate of Change
Emerging Technology Workshop, November 30, 2010, at NSF

http://www.nitrd.gov/et

Seeing Through Cloud Computing
Expedition Workshop #83, October 19, 2010, at the National Academy of Public Administration

History of Federally Funded Research and Development Centers (FFRDCs)
FASTER
April 27, 2010

Clifford A. Jacobs
National Science Foundation
Division of Atmospheric Sciences (GEO/ATM)

EarthCube - National Data Infrastructure for Earth System Science
Slides
June 19, 2012

Abstract

EarthCube is a new and novel approach taken by NSF to actively promote a series of ongoing engagements of multiple earth centered science(s) communities with relevant technologists. They include cyberinfrastructure developers, institutions and interested agencies contributing to create integrated infrastructure to transform the conduct of research within the geosciences for the purpose of accelerating our ability to understand and predict the Earth system. NSF created an intellectual commons where dialog and relationships can be nurtured with expressed intent to foster convergence of ideas and approaches for using the best technology solutions to advance science and enable substantial increases in productivity and capability of the community of researchers and educators.

EarthCube Community engagement site - http://earthcube.ning.com/
The goal of EarthCube is to transform the conduct of research by supporting the development of community-guided cyberinfrastructure to integrate data and information for knowledge management across the Geosciences. This website has been set up to foster community collaboration, and will provide updated information, resource documents, and
discussion forums so that community groups, consortia, researchers, and educators can share ideas, introduce concepts, and find and develop collaborative efforts.

**About The Speaker**

Dr. Clifford A. Jacobs

Expert  
National Science Foundation  
Directorate for Geosciences (GEO/OAD)  
703 292-4705  
cjacobs@nsf.gov  
http://www.nsf.gov/geo/earthcube/index.jsp

Dr. Jacobs is an NSF Expert. He retired from the National Science Foundation (NSF) after more than 25 years and provided oversight to the National Center for Atmospheric Research (NCAR) and activities at the University Corporation for Atmospheric Research (UCAR). His oversight responsibilities covered a wide range of topics ranging from supercomputers to aircraft and from climate modeling to impacts on society resulting from natural and anthropogenic induced changes in the environment. Dr. Jacobs represented geoscience in a variety of NSF studies and initiatives related to high performance computing and information technology, including the Blue Ribbon Panel on Cyberinfrastructure.

Prior to coming to NSF, Dr. Jacobs was executive VP and senior research scientist at The Center for the Environment and Man (CEM) in Hartford, CT. His basic research interests included four-dimensional computer models of the ocean, atmosphere and land processes, data analyses of large environmental databases, and the development of computer graphics software for the analysis of observed and model data. Domestic and foreign governments as well as private industry sponsored Dr. Jacobs’ research.

Dr. Jacobs received his Bachelor of Arts degree in Mathematics from Texas A&M University and his Master of Science degree in Oceanography, also from Texas A&M University. His Doctor of Philosophy degree was awarded by New York University in Oceanography.

---

**Data on Data**

**Presenting Stakeholder Alignment Data on the Cyberinfrastructure for Earth System Science**

Big Data SSG and FASTER

January 22, 2013

**Abstract**

Professor Cutcher-Gershenfeld will present mid-process findings from stakeholder alignment research on the NSF EarthCube initiative, assessing the social infrastructure for this leading “big data” cyberinfrastructure initiative. Based on survey data from over 850 geoscientists and computer scientists, the presentation will feature innovative methods for specifying stakeholders, identifying interests, and visually representing stakeholder alignment in complex systems.
Specific challenges will be identified in sharing data, tools, models, and visualizations across fields and disciplines; as well as underlying theory and principles for stakeholder alignment in complex systems. The research is supported through one of the inaugural NSF INSPIRE grants, bridging multiple directorates with transformational research (NSF EAR 12-49607).

EarthCube Community engagement site - http://earthcube.ning.com/
The goal of EarthCube is to transform the conduct of geoscience research through the development of a community-guided cyberinfrastructure, integrating data, tools, models, visualizations, and information across geoscience fields and disciplines. The EarthCube community website has been set up to foster community collaboration, and will provide updated information, resource documents, and discussion forums so that community groups, consortia, researchers, and educators can share ideas, introduce concepts, and find and develop collaborative efforts.

About The Speaker
Joel Cutcher-Gershenfeld
Professor, School of Labor and Employment Relations,
University of Illinois at Urbana-Champaign

Professor Cutcher-Gershenfeld’s research centers on enabling institutions to operate effectively in the context of accelerating change, expanding complexity, and increasing polarization. This includes advancing theory, policy, and practice associated with stakeholder alignment in complex systems. Professor Cutcher-Gershenfeld has co-authored six books, including “Strategic Negotiations” (Harvard Business School Press), “Knowledge-Driven Work” (Oxford University Press), “Lean Enterprise Value” (Palgrave), “Organizational Learning Systems” (Oxford University Press), and “International Human Resource Management and the Law” (Edwards Elgar). He is editor or co-editor of four additional books on human relations at work and workplace training, and the author or co-author of over 85 articles, book chapters and policy papers on negotiations, high performance work systems, labor-management relations, dispute resolution, economic development, and engineering systems. Professor Cutcher-Gershenfeld holds a Ph.D. from MIT and a B.S. from Cornell, both in industrial relations.
joelcg@illinois.edu

Presenting PAGES
(Public Access Gateway for Energy and Science)

FASTER
February 26, 2013

Abstract

Dr. Warnick will present PAGES (Public Access Gateway for Energy and Science), a web-based portal that will ensure that, after an embargo period, scholarly publications sponsored by the Department of Energy (DOE) are publicly accessible and searchable at no charge to readers. PAGES is designed to take advantage of the public access efforts of publishers by linking, via digital object identifiers (DOIs), to DOE articles they make publicly accessible. Each such article serves as the Version of Record, and it is hosted by the publisher. Thus, PAGES will avoid duplicating the public
access efforts of publishers. When DOE articles are not publicly accessible, PAGES will focus on accepted manuscripts. Specifically, after an embargo period, it will link, via URLs, to publicly accessible manuscripts hosted by institutional repositories. For those instances where free public access is offered neither by a publisher nor by an institutional repository, the DOE Office of Scientific and Technical Information will host the accepted manuscript and display it after an embargo period. In both of these cases, PAGES will still provide DOI links to publishers’ websites, where articles may be accessed with a subscription or other transaction, thus maintaining a pathway to the Version of Record.

Regardless of where DOE-sponsored articles or accepted manuscripts are hosted, PAGES will enable readers to search them all via a single search box. Among the metadata thus returned by PAGES will be the DOI for the published article once that DOI is posted by the publisher. PAGES will also be integrated with other DOE publicly-accessible R&D information and products, such as the 330,000 technical reports in the DOE Information Bridge.

**About The Speaker**

Walter L. Warnick, Ph.D.
Director, Office of Scientific and Technical Information (OSTI)
U.S. Department of Energy

Since becoming Director of the [Office of Scientific and Technical Information (OSTI)](https://semanticommunity.info/A_NITRD_Dashboard/Making_the_Most_of_Big_Data) in 1997, Dr. Warnick has championed an aggressive effort to capitalize on technological advances to provide state-of-the-art information tools and services to the DOE community, intergovernmental and international partners, students and the public. During his tenure at OSTI, groundbreaking tools have made DOE information access quicker, cheaper, more convenient, and more complete than ever before. For example, the DOE Science Accelerator provides one-stop searching of recent R&D results (including thousands of full-text technical reports and over one million e-print documents as well as science conference proceedings), R&D project accomplishments and other science sources.

New technology has been used to improve scientific and technical information management across the entire DOE complex. The Scientific and Technical Information Program, coordinated by OSTI, has successfully transitioned the DOE complex so that R&D results recorded in digital documents remain digital throughout their entire life cycle. Recognizing that the boundaries of science disciplines do not necessarily coincide with organizational boundaries, Dr. Warnick championed the interagency development of Science.gov, the gateway for R&D results and related government science information, and WorldWideScience.org, its international counterpart.

Dr. Warnick's vision for the future rests on the premise that emerging computing power and fast networks have only just begun to revolutionize scientific communication. "As the nation embarks on a new Science Information Infrastructure, text and communication tools will play a central role in ways not yet imagined," said Dr. Warnick. "But that's what OSTI is all about. As the information technology revolution evolves, we at OSTI will continue to fulfill our mandate to make DOE research information available. We are eager to help shape the future."

In 2005, Dr. Warnick was elected an American Association for the Advancement of Science (AAAS) Fellow in the Information, Computing, and Communication section “for leadership in the federal scientific information community and
for contributions to the conceptualization, development and implementation of innovative programs that significantly advance access to government information."

In April 2004, Dr. Warnick was chosen by the Public Printer of the United States to serve a three-year term on the 15-member Depository Library Council. "These individuals have dedicated their lives to ensuring that every American has access to Government information," said Public Printer Bruce James of the 2004 appointees. Also in 2004 Dr. Warnick was appointed chair of CENDI, an interagency working group of senior Scientific and Technical Information (STI) Managers from 13 U.S. federal agencies. CENDI is committed to addressing science- and technology-based national priorities and strengthening U.S. competitiveness.

In 2001, Dr. Warnick received both the Department of Energy Information Technology Quality Award for Executive Leadership and the government-wide IRMCO (Interagency Resources Management Conference) Individual Award for demonstration of exceptional ability to operate across organizational boundaries to improve the Government's service to its people.

Dr. Warnick obtained a bachelor's degree in engineering science from The Johns Hopkins University, and MS and PhD degrees in mechanical engineering from the University of Maryland.
walter.warnick@science.doe.gov

---

Cloud Computing and People with Disabilities

HCI&IM CG, SEW CG, and FASTER

Slides
March 26, 2013

Abstract

The rise of services supported "in the cloud", on the worldwide population of interconnected computers, is revolutionizing many businesses, while providing consumers with increased convenience at lower cost. What does this revolution in technology offer to people with disabilities? This talk will describe work being done in the USA and abroad to realize the Global Public Inclusive Infrastructure (GPII.net), using the cloud to make it much easier for people to access online content and services in a way that meets their individual needs and preferences. It will also outline further implications of the cloud for improvement of services for people with disabilities, through advances in "Big Data" analytics, in data sharing technology, and in social software.

About The Speaker

Clayton Lewis, Ph.D.
Professor of Computer Science and Fellow of the Institute of Cognitive Science
University of Colorado

Clayton Lewis is Professor of Computer Science and Fellow of the Institute of Cognitive Science at the University of Colorado. He is well known for his work (with students and colleagues) on evaluation methods in user interface design, including the thinking aloud and cognitive walkthrough methods. His recent work on technology for people with cognitive
Cloud & Accessibility Workshop

FASTER

June 12, 2013

Workshop Information

The NIST Cloud Computing Program will host a 1-day workshop on “Cloud & Accessibility” to investigate and discuss the relevant issues in this paradigm. The NIST Cloud Computing Roadmap (NIST SP 500-293) identifies the high priority requirements for complete adoption of cloud computing. Addressing accessibility falls under the realm of Requirement #7 of the Roadmap:

*Defined unique government regulatory requirements and solutions*

Ensuring that all citizens have equal access to information is a unique government regulatory requirement and accessibility is considered to be fundamental to this solution. Cloud computing solutions that address and highlight accessibility offer a path forward for an agency to fulfill its mission and requirements by providing a larger number of potential solutions that USG IT managers can use to be creative in the development of new services and solve their unique accessibility requirements. As work progresses in cloud computing, it is important to promote, incorporate and discuss applicable standards in accessibility for cloud computing services as a discipline for investigation.

The Workshop will discuss topics on Accessibility in the fundamental cloud computing application categories that are of importance to the USG:

- Enterprise collaboration Tools – email, calendaring, IM
- Resource/budget planning/management tools
- Customer relationship management (CRM)
- Web server /Content Management
- Identity Management
- Document Retrieval/ Library Systems

The workshop will also host an “Industry Response” and “Path Forward” panels and scheduled breakouts in the afternoon. The workshop agenda and list of speakers will be posted shortly at [http://www.nitrd.gov/rsvp/cloud_workshop.aspx](http://www.nitrd.gov/rsvp/cloud_workshop.aspx)
Abstract

ITU-T Recommendation X.1255 provides an open-architecture framework for discovery of identity management information. This Recommendation began as a contribution to ITU-T Study Group 17 for a standard on interoperability between heterogeneous information systems. Although there was considerable interest in the contribution at the time, it was assigned to a working group on identity management, which quickly determined the most pressing aspect of their work was on discovery of such information. So, the contribution and the language therein were redirected to that topic.

After much discussion and commentary within the working group, it was concluded that the specific suggestion be made somewhat more abstract so that multiple contributions would be possible for components of the proposed Recommendation. The result was a framework document that suggests how to think about achieving the desired end result of discovery (or interoperability more generally), without specifying exactly how to go about it. This is analogous to describing the Internet by starting with specific protocols to implement and then abstracting those protocols into a framework of generic objectives.

The approach to discovery of identity management information is not really different from discovery of any other kind of information in digital form. It requires the ability of individual systems to share information of their choice in a federation of such systems. Both hierarchical and peer-peer federations are discussed along with the means by which independent systems can exchange information with administrative controls.

The Recommendation was based on technology developed by Dr. Robert Kahn and his team at Corporation for National Research Initiatives® (CNRI) in Reston, Virginia. Dr. Kahn is a co-inventor of the TCP/IP protocols and was responsible for originating DARPA’s Internet Program.

About The Speaker

Robert E. Kahn, Ph.D.
Chairman, CEO and President of the Corporation for National Research Initiatives (CNRI)

Robert E. Kahn is Chairman, CEO and President of the Corporation for National Research Initiatives (CNRI), which he founded in 1986 after a thirteen year term at the U.S. Defense Advanced Research Projects Agency (DARPA). CNRI was created as a not-for-profit organization to provide leadership and funding for research and development of the National Information Infrastructure.

After receiving a B.E.E. from the City College of New York in 1960, Dr. Kahn earned M.A. and Ph.D. degrees from Princeton University in 1962 and 1964 respectively. He worked on the Technical Staff at Bell Laboratories and then
became an Assistant Professor of Electrical Engineering at MIT. He took a leave of absence from MIT to join Bolt Beranek and Newman, where he was responsible for the system design of the Arpanet, the first packet-switched network. In 1972 he moved to DARPA and subsequently became Director of DARPA's Information Processing Techniques Office (IPTO). While Director of IPTO he initiated the United States government's billion dollar Strategic Computing Program, the largest computer research and development program ever undertaken by the federal government. Dr. Kahn conceived the idea of open-architecture networking. He is a co-inventor of the TCP/IP protocols and was responsible for originating DARPA's Internet Program. Dr. Kahn also coined the term National Information Infrastructure (NII) in the mid-1980s which later became more widely known as the Information Super Highway.

In his recent work, Dr. Kahn has been developing the concept of a digital object architecture as a key middleware component of the NII. This notion is providing a framework for interoperability of heterogeneous information systems and is being used in many applications such as the Digital Object Identifier (DOI). He is a co-inventor of Knowbot programs, mobile software agents in the network environment.

Dr. Kahn is a member of the National Academy of Engineering, a Fellow of the IEEE, a Fellow of AAAI, a Fellow of ACM and a Fellow of the Computer History Museum. He is a member of the State Department's Advisory Committee on International Communications and Information Policy, and the National Archives & Records Administration Advisory Committee on the Electronic Records Archives (ACERA). Among his many awards, he is a recipient of the National Medal of Technology, ACM Turing Award, Queen Elizabeth Prize for Engineering and the Presidential Medal of Freedom.

Cloud Service Level Agreements

Meeting Customer and Provider Needs

Slides

FASTER

January 28, 2014

Abstract

Purchasing cloud computing services is a hazy process and for the Federal Government there are additional requirements which make this process even more challenging. To correctly acquire cloud services, one must be able to understand the business requirements, match the requirements to a provider's cloud service offering, reach an agreement with a provider on what will be provided, and measure the performance of the service.

Mr. Simmon will discuss the NIST activities in developing a customer decision making framework, SLA standards, and cloud metrology to assist both the customer and provider in achieving clarity when procuring and using cloud computing services.

About The Speaker

Eric Simmon
NIST Cloud Computing Program

https://semanticommunity.info/A_NITRD_Dashboard/Making_the_Most_of_Big_Data
Updated: Thu, 27 Jun 2019 05:55:32 GMT
Powered by mindtouch
 Eric Simmon is a systems expert in the Electronic Information Group (EIG) at the National Institute of Standards of Technology working on using systems modeling to improve the standards development process. He is chair of the Federal Cloud Computing Standards and Technology Working Group and project editor for the ISO/IEC 19086 Cloud Service Level Agreement standardization effort.

Opening Government Data and Creating Public Value

FASTER, and the Big Data SSG

Slides: I HAVE THESE

April 22, 2014

Abstract

When opening government data a common assumption is that simply supplying more data freely and in more formats will lead to use. Use will then lead to value creation, and in turn, governments will be motivated to make the necessary changes to continue to open more data. Experience shows that supplying more and more data does not necessarily produce the results we anticipated. This talk will highlight work carried out by the Center for Technology in Government at the University at Albany, SUNY, together with government entities around the world, to examine the range of public value that accrues to different stakeholders when opening and using government data. The Center’s research to practice translation model will be discussed.

About The Speaker

Dr. Theresa A. Pardo
Director, Center for Technology in Government, University at Albany

Dr. Theresa A. Pardo serves as Director of the Center for Technology in Government at the University at Albany, State University of New York, where she also holds research professor appointments in Public Administration and Policy and Informatics. Dr. Pardo is co-developer of the top ranked program in Government Information Strategy and Management offered by the Public Administration and Policy Department of Rockefeller College.

Under her leadership, the Center works closely with multi-sector and multi-disciplinary teams from the U.S. and around the world to carry out applied research and problem solving projects focused on the intersections of policy, management, and technology in the governmental context. With funding from organizations such as the U.S. National Science Foundation, Ford Foundation, U.S. Department of Justice, United Nations, World Bank, U.S. Library of Congress, U.S. National Archives and Records Administration, SAP, Microsoft Corporation, and New York State, among others, Center work has broken ground in information and knowledge sharing, open government and open data, e-government, social media policy, and mobile technologies and human services delivery.

Dr. Pardo serves as Open NY Adviser to New York State’s Governor Andrew Cuomo and was recently appointed Expert Advisor to the Open Data 500 Project. She serves as International Advisor to the E-Government Committee for the China Information Association and is President of the Digital Government Society, a global multi-disciplinary organization of scholars and practitioners interested in the development and impacts of digital government. Dr. Pardo
serves on a number of national and international advisory and editorial boards, including Government Information Quarterly, the leading journal in the field of digital government, the U.S. Government Accountability Office Executive Council on Information, Management, and Technology, and the Data Center for Applied Research in Social Sciences at Centro de Investigacion y Docencia Economicas (CIDE) in Mexico City. She is also a Senior Adviser to the Informatization Research Institution, State Information Center in China and is a member of the Board of Champions for the NYS STEAM Girls Collaborative.

Dr. Pardo has published over 125 articles, research reports, practice guides, book chapters and case studies. She has received numerous awards including best journal and conference paper awards for her published work, the University at Albany’s Excellence in Teaching Award and the Rockefeller College Distinguished Service Award.

Update on CASC HPC & Cyberinfrastructure Developments

FASTER

May 27, 2014

Abstract

An informative discussion with Sue Fratkin who will share her perspectives regarding the Coalition for Academic Scientific Computation (CASC) High Performance Computing (HPC) and cyberinfrastructure developments. Mrs. Fratkin serves as the Washington liaison for the CASC, an alliance of forty-one academic supercomputing centers in twenty-eight states.

About The Speaker

Sue Fratkin
Fratkin Associates (owner), and Washington Liaison for the Coalition for Academic Scientific Computation (CASC)

Sue Fratkin is a public policy analyst concentrating on technology and telecommunications issues, particularly as they affect the higher education community. Realizing the impact that the growth in technology would have on this community, Sue founded Fratkin Associates in 1991. She regularly interacts with White House, Congressional and Federal Agency personnel and reports on governmental hearings, programs and publications, analyzing relevant legislation and regulations.

Fratkin Associates’ clients include several Washington based higher education and IT focused associations. She also serves as the Washington liaison for the Coalition for Academic Scientific Computation, an alliance of forty-one academic supercomputing centers in twenty-eight states.

As the author of numerous articles on public policy and telecommunications and technology policy pertaining to the higher education community, Sue participates as a panelist at national conferences and has served as a reviewer for Federal education technology programs.
Workshop Information

As part of procuring cloud services from a cloud provider, a cloud customer must decide exactly what types of services are needed, a way to ensure that these services are genuinely provided and accountable, and also a contract mechanism to state the level of performance and other characteristics which pertain to the cloud services. The NIST Cloud Computing Standards & Technology Roadmap (NIST SP 500-292) identified a specific Requirement for: Technical specifications to enable development of consistent, high-quality Service-Level Agreements. To fully understand the scope of contracts and service level agreements in cloud computing, it is necessary to examine the definitions of each:

Service agreement: A legal document specifying the rules of the legal contract between the cloud user and the cloud provider. (NIST SP 800-146)

Service-level agreement: A document stating the technical performance promises made by the cloud provider, how disputes are to be discovered and handled, and any remedies for performance failures. (NIST SP 800-146)

This workshop will explore avenues to assist cloud customer and provider in the procurement of cloud services by determining methods for constructing high-quality SLAs.

Specifically, the workshop (and breakout sessions) will:

• Explain the concepts behind SLAs
• Describe current work on standardizing the building blocks of SLAs (concepts, terms, definitions and contexts)
• Discuss Services in terms of Performance, Management & Security
• Show the relationship of cloud service metrics and SLAs

The workshop agenda and list of speakers are posted at: http://www.nitrd.gov/nitrdgroups/index.php?title=Cloud_SLA_Workshop

Data Infrastructure Building Blocks (DIBBs)

FASTER

Slides

August 19, 2014
Abstract

The Data Infrastructure Building Blocks (DIBBs) program is an integral part of Cyberinfrastructure Framework for 21st Century Science and Engineering (CIF21). The DIBBs program encourages development of robust and shared data-centric cyberinfrastructure capabilities to accelerate interdisciplinary and collaborative research in areas of inquiry stimulated by data.

Effective solutions will bring together cyberinfrastructure expertise and domain researchers, to ensure that the resulting cyberinfrastructure components address the researchers' data needs. The activities address the data challenges arising in a disciplinary or cross-disciplinary context.

About The Speaker

Alexander Szalay
Alumni Centennial Professor
Dept. of Physics and Astronomy - The Johns Hopkins University

Alexander Szalay is the Alumni Centennial Professor of Astronomy at the Johns Hopkins University. He is also a professor in the Department of Computer Science. A cosmologist, he works on the statistical measures of the spatial distribution of galaxies and galaxy formation.

He was born and educated in Hungary, and he spent postdoctoral periods at UC Berkeley and the University of Chicago before accepting a faculty position at Johns Hopkins. Szalay was elected to the Hungarian Academy of Sciences as a corresponding member in 1990.

He is the architect for the Science Archive of the Sloan Digital Sky Survey and project director of the NSF-funded National Virtual Observatory. He has written more than 340 papers that have appeared in various scientific journals, covering areas such as theoretical cosmology, observational astronomy, spatial statistics, and computer science.

In 2003 he was elected as a Fellow of the American Academy of Arts and Sciences. He received an Alexander von Humboldt Prize in Physical Sciences in 2004 and a Microsoft Award for Technical Computing in 2008. He is a Corresponding Member of the Hungarian Academy of Sciences, and a Fellow of the American Academy of Arts and Sciences. In 2008 he became Doctor Honoris Clausa of the Eotvos University.

NEXT

Semantic Data Science Team Attends White House Big Data Event

Tweet: #Data2Action Really enjoyed our Semantic Data Science Team participating: http://semanticommunity.info/A_NITRD_Dashboard/Making_the_Most_of_Big_Data#Story …

Our Semantic Data Science Team has been invited to attend the event on November 12th that is described as follows:
The unveiling of new, high-impact collaborations among Federal agencies, private industry, academia, state and local governments, non-profits, and foundations, that will harness Big Data to enhance economic growth and job creation, education and health, energy and environmental sustainability, public safety, and global development.

Our Team has been working towards this since our presentation to the Federal Big Data Senior Steering Work Group last January and our most recent Cloud, SOA, Semantics and Data Science Conference in September where we presented our Semantic Medline with YarcData Graph Appliance Pilot: Wiki YarcData Videos (Schizo-7 minutes, Cancer-21 minutes).

Our work is an example of the bold new collaboration theme: “Harnessing the Potential of Data Scientists and Big Data for Scientific Discovery” that shows “Data Innovation Across Sectors” and includes the following Breakout session topics:

- Education and Workforce Development (George Mason University and John Hopkins University - see below)
- Research and Development (NIH and YarcData)
- Innovation (DC Data Science Community and Semantic Community)

We have submitted a two-page summary on Making the Most of Big Data to BigDataprojects@nitrd.gov and indicated that NITRD can post the project description to a public website.

About a year ago, Dr. George Strawn challenged me to pilot a new partnership to make NIH’s Semantic Medline “the killer semantic web app for the government” and the keystone of workforce development for future data scientists.

I formed a Data Science Team from NIH, YarcData, George Mason University, and my Semantic Community to migrate Semantic Medline to the new YarcData Graph Appliance for improved speed and scalability and to help start a Data Science program at GMU where students would use it for research papers and dissertations.

The demo videos are now on YouTube (see above), more details are available at the Web Site we prepared for NITRD (this Wiki Page), and we have a Fact Sheet: “Accelerating Disease Research on NIH’s MEDLINE using a “Big Data” Approach” (DRAFT)

We are very pleased with our progress and we are in discussions with JHU Genetics and the Cancer Genome Atlas (TCGA) about using their data sets next.

We have worked to support the NITRD Current and Planned Coordination Activities as follows:

- Working with two of the six agencies: NSF, NIH, and trying to work with the other four: DoD, DARPA, DOE, and USGS;
- Following the work in the NSF-NIH Solicitation, Core Techniques and Technologies for Advancing Big Data Science & Engineering for datasets and results that can be reused;
- Helping ensure a trained workforce to capitalize on big data resources by working with GMU Data Science as part of our team and preparing a graduate course on data science using the applications and data sets mentioned above and below;
- Providing examples of applications that use multiagency big datasets and core technology that is needed to turn heterogeneous data into more homogeneous, interoperable data;
- Providing big data infrastructure development for domain science with Spotfire and the YarcData Graph Appliance; and

https://semanticommunity.info/A_NITRD_Dashboard/Making_the_Most_of_Big_Data
Updated: Thu, 27 Jun 2019 05:55:32 GMT
Powered by mindtouch
• Attending the second National Big Data R&D Initiative event.

Our Semantic Data Science Team has done extensive data science work for OMB, Japan, EU, Data.gov etc. which is reported at: http://semanticommunity.info/

We are planning a series of webinars and a conference on the use of Be Informed in the US with major applications as follows:

• HealthCare.gov (Like the Be Informed Permit Application)
• LDS Church: (Like the Be Informed Process-driven Documentation Application)
• NIH Semantic Medline: (Like Be Informed's Leveraging Big Data Analytics in Business Processes)

This event was titled "Building New Partnerships", and the next event should be titled "Building New Applications" like our "Accelerating Disease.Research on NIH's MEDLINE using a "Big Data" Approach. (DRAFT)

MORE SPECIFICS TO FOLLOW BASED ON MINING THE TECHNICAL SHOWCASE AND FACT SHEETS

JHU Genetics

After watching the videos, I wrote the following summary for our use:

SemRep AND SemMed

SemRep <http://semrep.nlm.nih.gov>, is an open source Java application that uses natural language processing techniques to divide texts from PubMed's Medline citation database <http://www.ncbi.nlm.nih.gov/pubmed/> into "sentences" and parse them in order to extract RDF triples and predications from them - semantically tightly-coupled subjects, objects, and verbs (predicates). It then normalizes the vocabulary of the predications using the standards Unified Medical Language System and Entrez Gene as base ontologies.


* summarizes it by, among other things, removing non-salient triples;
* creates an RDF triple store graph database from the resulting salient triples;
* provides a Java EE-based web application for querying the graph database; and
* presents query results visually as directed graphs that can be browsed and re-queried.

SIMPLIFIED SemRep/SemMed WORKFLOW

When SemRep is fed a Medline citation of an article titled "Exemestane after non-steroidal aromatase inhibitors for post-menopausal women with advanced breast cancer", it will extract, among others, the following predication from the title:

"aromatase inhibitors" "[are] for" "breast cancer"

SemRep then normalizes the predication to UMLS standards and generates the following [pseudo] RDF triple:
SemRep has thus extracted an element of meaning from the text and put it into a computable form (an RDF triple) that can be stored into a triple store graph database and queried. (Each element of the RDF triple contains a URI link to its source text, allowing the user to read the actual source article if desired.)

SemRep will also extract another predication from the title:

[BREAST_CARCINOMA] -> [PROCESS_OF] -> [INDIVIDUAL]

But this predication, because it offers no salient information in our context, will be eliminated when SemMed "summarizes" SemRep’s output prior to database creation.

STATISTICS

SemRep has already parsed 22.5 million Medline citations into 65.5 million RDF predications derived from 2.2 billion RDF triples. (I’m guessing the run-time RAM footprint for that much data is about 700gB.) The Cray YarcData Urika computer appliance allows all of this data to be in memory at once, thereby eliminating the potentially enormous performance hit that would come from sharding this large graph database.

This summary of SemRep and SemMed is based partially on information gleaned from the above-mentioned websites and from the following videos:

MEDLine Schizophrenia Complete Demo
<http://www.youtube.com/watch?v=Shfl4SNzNO4>

MEDLine Cancer Complete Demo
<http://www.youtube.com/watch?v=6frNAmPD0mo>

It would be interesting to me to know how much RAM this database uses and how fast the queries execute.

I am looking forward to testing Semantic Medline soon.

Respectfully,

Dean A. Snyder

Senior Programmer/Analyst

Center for Inherited Disease Research (CIDR)

McKusick-Nathans Institute of Genetic Medicine (IGM)

Johns Hopkins University School of Medicine

333 Cassell Drive, Triad Technology Ctr., Suite 2000
Linked TCGA: A clinically relevant 30 billion triple Dataset


Cambridge Semantic Web Monthly Meetup

Tuesday, November 12, 2013, 6:00 PM

[MIT Stata Center - Star Room](https://www.mit.edu/maps/)

32 Vassar Street
Star Room on 4th Floor
Cambridge, MA (map)

Rm 32-D463, 4th floor, Star Conference Room, Stata Center, MIT

Agenda:

Helena Deus, Senior Scientist and Medical Knowledge Engineer at Foundation Medicine ()

https://semanticommunity.info/A_NITRD_Dashboard/Making_the_Most_of_Big_Data
Updated: Thu, 27 Jun 2019 05:55:32 GMT
The Cancer Genome Atlas (TCGA) is a multidisciplinary, multi-institutional effort to catalogue genetic mutations responsible for cancer using genome analysis techniques. It consists of over 400Tb of valuable clinically actionable data which doubles in size every 9 months.

In this talk I will describe how we built TopFed, our robust and distributed solution for supporting clinically driven SPARQL queries.

Brand Niemann

Can we collaborate with you to try your 30B Triples in the YarcData Graph Application? We have had success with Semantic Medline: http://semanticommunity.info/A ...
We have asked Kingsley to help us.

Data to Knowledge to Action Event Highlights Innovative Collaborations to Benefit Americans

Corporations, Universities, Philanthropies, Nonprofits, and State and Local Governments Answer President’s Call for New Partnerships

Source: November 12, 2013 Email

FOR IMMEDIATE RELEASE

November 12, 2013

Contacts:

Rick Weiss: 202 456-6037: rweiss@ostp.eop.gov
Moira Vahey: 202 456-3545: mvahey@ostp.eop.gov

Dozens of public and private organizations are meeting at a White House-sponsored event today to describe their contributions to an inspiring array of collaborations that embrace a common theme: sharing resources and drawing upon sophisticated tools to plumb the depths of huge data sets in order to derive greater value for American consumers and grow the Nation’s economy.

The event, at the Ronald Reagan Building and International Trade Center in Washington, DC, is co-sponsored by the White House Office of Science and Technology Policy and the Networking and Information Technology R&D (NITRD) program, which represents the information technology portfolios of 18 Federal agencies. It features scores of new announcements by corporations, educational institutions, professional organizations, and others that—in collaboration with Federal departments and agencies and state and local governments—are answering President Obama’s call for partnerships that can enhance national priorities, including economic growth and job creation, education and health, energy and sustainability, public safety and national security, and global development.
“America is rich with institutions that are expert at generating data, but as a Nation we have not fulfilled our potential to make the most of these data by merging pre-competitive resources, partnering on analytics, and sharing lessons learned,” said John P. Holdren, Assistant to the President for Science and Technology and Director of the White House Office of Science and Technology Policy. “Today’s announcements show that we are maturing in this respect, finding synergies and collaborative opportunities that will accelerate progress in a wide range of scientific, social, and economic domains.”

The new commitments build on an initial platform of projects launched last year as part of the Obama Administration’s $200 million “Big Data Initiative.” Among the new announcements:

**Building virtual clinical trials**

Tremendous progress against cancer has been realized through clinical trials, but only a small proportion of all cancer patients participate in such trials. The American Society of Clinical Oncology aims to realize some of the potential from thousands of other cancer patients through CancerLinQ, a five-year, $80 million initiative being undertaken with support from its Conquer Cancer Foundation and with donations from Amgen, Chan Soon-Shiong Family Foundation, Genentech BioOncology, Helsinn Therapeutics (US) Inc., Susan G. Komen, and numerous individual supporters. This learning computer network will unlock and analyze vast quantities of de-identified information on patient experiences that are now lost to file cabinets and unconnected servers, to help inform and guide medical professionals and patients.

**Improving access to clinical trials**

More than half of all patients say they are interested in participating in clinical trials yet nearly half of all clinical trials fail to reach their recruitment targets. To help close this gap, Novartis, Pfizer, and Eli Lilly and Co. are partnering in the United States to provide a new platform to improve patient access to information about clinical trials. The platform will enhance the existing clinicaltrials.gov clearinghouse website by providing more detailed and patient-friendly information about available trials and embedding a machine-readable “target health profile” to improve the ability of healthcare software to match individual health profiles with applicable clinical trials.

**Bringing NASA data down to Earth**

Amazon Web Services (AWS) and NASA are making space-based data about the Earth widely available to the public through the NASA Earth eXchange (NEX), a collaborative sharing network for researchers in Earth Science. AWS is working with NEX to host a significant amount of NASA’s Earth-observing data as an AWS Public Data Set. Among other benefits, this will bolster projects like Citizen Science Alliance’s Zooniverse.org, which allows researchers to leverage the power of the crowd to quickly analyze massive data sets and work on problems that cannot be efficiently solved by computers. Already, for example, the Zooniverse’s Galaxy Zoo projects have used contributions from volunteer “citizen scientists” to classify more that 1 million galaxies in the Sloan Digital Sky Survey.
Applying data analytics to society’s problem-solvers

DataKind, a non-profit that matches data scientists with non-profit and non-governmental organizations, is partnering with Pivotal to bring some of industry’s top data analytics talent to bear on some of society’s greatest challenges. Many high-impact social organizations have huge troves of data but lack the resources to analyze them. Through its new data philanthropy initiatives, Pivotal’s data scientists can volunteer their skills and engage data scientists around the world. DataKind is also partnering with The Mission Continues to help that organization better understand the effects of its volunteer programs that aim to improve veterans’ lives. DataKind is also starting a new project with Medic Mobile—a non-profit that uses communication technologies to improve the health of under-served and disconnected communities—to better measure the impact of that organization’s various health initiatives.

A number of other organizations also announced new collaborations, including:

- NIH, IBM, Sutter Health, and Geisigner Health System, developing new methods for early detection of heart disease;
- NIH’s Big Data to Knowledge initiative, developing new standards and tools to help biomedical scientists use the big data being generated by the research community;
- New York University, University of California at Berkeley, and the University of Washington, together harnessing big data for basic research and scientific discovery, with support from the Gordon and Betty Moore and Alfred P. Sloan Foundations;
- DC-NET and CAAREN, partnering with George Washington University Medical Center on next-generation genomic sequencing;
- OSTP, forming an initiative to leverage big data to predict pandemics;
- City of New York, helping new businesses get off the ground more quickly using data across city agencies;
- The Kamusi Project, aligning with universities around the world and “citizen linguists” to create a dictionary of every word in every language;
- MIT and the city of Boston, launching a big-data challenge on urban transportation;
- Splunk, NoticeandComment, and Sunlight Foundation, facilitating Federal and local civic engagement using Regulations.gov;
- Berkeley’s AMPLab, funded by NSF’s Expeditions in Computing program, is releasing an open-source big data analytics stack for lightning-fast cluster computing that can be used in large-scale collaborations;
- IBM, partnering with universities and industry to create a tool to help position students for the most in-demand data jobs;
- MIT, forming a new working group on big data and privacy; and
- A new Council for Big Data, Ethics, and Society, which will provide critical social and perspectives on big data initiatives.

Suzi Iacono, co-chair of NITRD’s Big Data Senior Steering Group and Deputy Assistant Director of the National Science Foundation’s Directorate for Computer and Information Science and Engineering, said she was gratified by the tremendous response to the Administration’s call for collaborative efforts in the domain of data innovation.

“We are seeing progress on so many fronts, for example developing the foundations of scalable algorithms, integrating human and machine reasoning in large-scale inferences, extracting knowledge from large, diverse, and complex data sets, and altogether facilitating powerful new approaches to discovery and decision-making,” Iacono said. “With a
sustained commitment from the Administration and our many partners, novel scientific discoveries and wide-spread innovation are certainly on the horizon.”

- For a full list of commitments presented at today’s event, please click HERE (PDF)
- For a fact sheet on Data to Knowledge to Action: Progress by Federal Agencies, please click HERE (PDF)
- For more information about OSTP, please visit: http://ww.whitehouse.gov/ostp
- For more information about NITRD, please click HERE (PDF)

Becky Fried

Communications Analyst

The White House | Office of Science & Technology Policy

O: 202.456.6045

M: 202.503.7122

E: rfried@ostp.eop.gov

Full list of commitments presented at today’s event
please click HERE (PDF)

Fact sheet on Data to Knowledge to Action: Progress by Federal Agencies
please click HERE (PDF)

More information about NITRD
please click HERE

NITRD Supplement to the President’s Budget

Source: http://www.whitehouse.gov/sites/default/files/microsites/ostp/nitrd_fy14_budgetsup.pdf (PDF)

Additional Program Focus Areas
Big Data R&D (BD R&D)

Participating Agencies: DARPA, DoD Service Research Organizations, DOE/NNSA, DOE/SC, EPA, NASA, NIH, NIST, NOAA, NRO, NSA, NSF, OSD, Treasury/OFR, USAID, and USGS

The Big Data Senior Steering Group (BD SSG) was formed in early 2011 to identify current big data research and development activities across the Federal Government; offer opportunities for coordination between agencies, academia, and the private sector; and identify the goals for a National Big Data R&D Initiative. As data volumes grow exponentially, so does the concern regarding data preservation, access, dissemination, and usability. Research in areas such as database interoperability, machine learning, automated analysis, visualization, and improved data privacy and security will help advance science and inspire revolutionary research techniques.
Strategic Priorities
The BD SSG envisions a future in which the ability to analyze and extract information from large, diverse, and disparate data sets accelerates the process of scientific discovery and innovation; promotes new economic growth; and leads to new fields of research and new areas of inquiry that would otherwise be impossible. The BD SSG strategic priorities include:

- Core technologies
- Big data infrastructure
- Workforce development
- Competitions and challenges

Current and Planned Coordination Activities
The current and planned coordination activities of the BD SSG include:

- In March 2012, the White House Office of Science and Technology Policy launched the National Big Data R&D Initiative. Six agencies – NSF, NIH, DoD, DARPA, DOE, and USGS – announced new big data projects at the launch event, including the joint NSF-NIH solicitation, Core Techniques and Technologies for Advancing Big Data Science & Engineering. The Fact Sheet: Big Data Across the Federal Government, released at the event, included a compilation of information provided by the BD SSG on over 65 federal projects.

- The joint NSF-NIH Solicitation, Core Techniques and Technologies for Advancing Big Data Science & Engineering, developed under the aegis of the BD SSG, is intended to “advance the core scientific and technological means of managing, analyzing, visualizing, and extracting useful information” from big data, particularly those technologies that are relevant across multiple agency missions. The solicitation generated great interest and resulted in several hundred high quality proposals. In addition to the mid-scale projects awarded in Fall 2012, small project awards are to be announced in Spring 2013.

- To help ensure a trained workforce to capitalize on big data resources, the BD SSG created an inventory of the workforce development programs at nine agencies. The group plans to use this information, combined with information gathered from academia and the private sector, to develop recommendations on how to develop an effective big data workforce.

- To explore the use of challenges and competitions, NSF, NASA, and DOE/SC jointly launched a series of Big Data Challenges beginning in October 2012. Receipt of submissions, judging, and awards started in November 2012 on a rolling basis. This series of challenges is designed not only to identify unique applications for the use of multiagency big datasets, but also as a way to identify the core technology that is needed to turn heterogeneous data into more homogeneous, interoperable data. The series is expected to continue throughout 2013.

- The BD SSG is examining big data infrastructure development for domain science (also known as domain research projects). The group identified four existing federal projects that involve big data and are, or could become, relevant to multiple agencies. Over the upcoming year, activities include exploring each project through interviews and joint meetings to determine the infrastructure needed to realize the full potential of currently available data or future data. The group plans to synthesize and use the information to develop recommendations for big data infrastructure.

- A second National Big Data R&D Initiative event is planned for 2013 to bring together federal, academic, and private sector stakeholders to discuss progress made and identify areas for investment.

11 [http://www.whitehouse.gov/sites/defa...se_final_2.pdf](http://www.whitehouse.gov/sites/defa...se_final_2.pdf)
12 [http://www.whitehouse.gov/sites/defa...heet_final.pdf](http://www.whitehouse.gov/sites/defa...heet_final.pdf)
13 Census, CIA, DHS, DoD, DOE, ED, NIH, NSA, NSF
14 [http://community.topcoder.com/coeci/nitrd](http://community.topcoder.com/coeci/nitrd)
15 Climate Change and Public Health; Materials Genome Initiative; Space Weather; Electronic Health Records
Data to Knowledge to Action Event Takes Big Data Initiatives to Innovative Heights

Source: October 7, 2013 Email

Following up on last year’s Obama Administration announcement of a “Big Data” initiative, the Office of Science and Technology Policy, National Science Foundation, and other agencies in the Networking and Information Technology Research & Development program will unveil new, high-impact collaborations among Federal agencies, private industry, academia, state and local governments, non-profits, and foundations, that will harness Big Data to enhance economic growth and job creation, education and health, energy and environmental sustainability, public safety, and global development.

DATE: Tuesday, November 12, 2013

TIME: 1:00 pm to 3:30 pm EST

LOCATION: Pavilion Room at the Ronald Reagan Building and International Trade Center, 1300 Pennsylvania Avenue, NW, Washington, DC

AGENDA: Available here.

This event will also be Webcast live at: http://live.science360.gov

Registration is not required for this event.

Becky Fried
Communications Analyst
The White House | Office of Science & Technology Policy

O: 202.456.6045
M: 202.503.7122
E: rfried@ostp.eop.gov

Building New Partnerships

1245 Coffee served in the Pavilion Foyer

(ALL SESSIONS AFTER 330 ARE BY INVITATION ONLY AND CLOSED TO THE PRESS)

100 Welcome
John P. Holdren
Assistant to the President for Science and Technology
Director of the White House Office of Science and Technology Policy

Farnam Jahanian
Co-chair of the Subcommittee for Networking & Information Technology R&D
Assistant Director for the CISE Directorate, National Science Foundation

115 Building New Partnerships

Thomas Kalil
Deputy Director for Technology and Innovation
Office of Science and Technology Policy

David Logsdon
Senior Director for Federal Civil, Global Public Sector
TechAmerica

200 A Bold New Collaboration

“Harnessing the Potential of Data Scientists and Big Data for Scientific Discovery”
Chris Mentzel
Program Officer, Data-Driven Discovery
Gordon and Betty Moore Foundation

Joshua Greenberg
Program Director, Digital Information Technology
The Alfred P. Sloan Foundation

Ed Lazowska
Bill & Melinda Gates Chair in Computer Science & Engineering
Director of the eScience Institute
University of Washington

Yann LeCun
Silver Professor of Computer Science and Neural Science,
Courant Institute of Mathematical Sciences, and Director of the Center for Data Science
New York University

Saul Perlmutter
Director of the Berkeley Institute for Data Science
Professor of Physics, University of California, Berkeley
Senior Scientist, Lawrence Berkeley National Laboratory

220 Coffee served in the Pavilion Foyer

230 Partnership Panel Discussion

“Data Innovation Across Sectors”
Marjorie Censer (Moderator)
Washington Post

Jake Porway
330 Coffee served in the Meridian Foyer

345 Breakout Sessions

**Education and Workforce Development**

**MERIDIAN ROOM B**
To discuss collaborative projects that focus on either improving the data workforce or educating the general public about data science and engineering. These could include: undergraduate and graduate curriculum and practicum; training and certification programs; K-12 education; informal Big Data and data analytics education; and improving access to the Big Data talent pool for organizations needing data scientists

Michael Rappa (Moderator)
Director, Institute for Advanced Analytics
North Carolina State University

Deborah Nolan
Professor, Department of Statistics
University of California, Berkeley

Roy Campbell
Director of Graduate Admission and Advancement
Sohaib and Sara Abbasi Professor
University of Illinois

**Research and Development**

**MERIDIAN ROOM DE**
To discuss new research partnerships across federal, academic, private, and non-profit sectors. Topics could include: core techniques and technologies; scientific and technical computing; cross-sector research centers; and cyberinfrastructure to support data science and engineering.
Innovation

MERIDIAN ROOM C
To discuss how the data community can come together to foster the innovation ecosystem and promote the use of data science and engineering to address the challenges our nation faces. These challenges could include: national and regional economic development, education, healthcare, energy, climate change, transportation, and others.

Christy Wilson (Moderator)
Vice President of Product Operations
Splunk

Eng Lim Goh
Chief Technology Officer
SGI

Joris VanDam
Strategic Projects Leader, Pharmaceutical Development
Novartis

445 Coffee served in the Meridian Foyer

500 Breakout Sessions Report

520 Closing

George Strawn
Director of the National Coordination Office
Networking and Information Technology R&D Program

Story

Slides and Wiki YarcData Videos(Schizo-7 minutes, Cancer-21 minutes)

Making the Most of Big Data

The Obama Administration announced a "National Big Data Research and Development Initiative" on March 29, 2012, to greatly improve and develop the tools, techniques, and human capital needed to move from data to knowledge to action.
Now in the second year of the Big Data Initiative, the Administration is encouraging multiple stakeholders to develop and participate in Big Data innovation projects across the country. Later this year, an event that highlights high-impact collaborations and identifies areas for expanded collaboration between the public and private sectors is planned. The Administration is particularly interested in projects and initiatives that:

- Advance technologies that support Big Data and data analytics;
- Educate and expand the Big Data workforce;
- Develop, demonstrate and evaluate applications of Big Data that improve key outcomes in economic growth, job creation, education, health, energy, sustainability, public safety, advanced manufacturing, science and engineering, and global development;
- Demonstrate the role that prizes and challenges can play in deriving new insights from Big Data; and
- Foster regional innovation.

Please submit a two-page summary of projects to BigDataprojects@nitrd.gov and indicate whether NITRD can post the project description to a public website.

The summary should identify:

1. The goal of the project, with metrics for evaluating the success or failure of the project;

Our Data Science Team had a very successful conference last week featuring the live joint demonstration of the NIH Semantic Medline utilizing the YARCDa Appliiance that was originally presented to the Federal Big Data Senior Steering Work Group (FBDSSWG) January 2013, following our kickoff presentation in the Fall of 2012.

We are being invited back to present to the FBDSSWG again and want to show the Semantic Medline – YARCData demo. We would also describe our efforts to involve George Mason University (GMU) in using this platform for student projects/dissertations in support of the Big Data Work Force Training Initiative.

Our first metric was to implement Semantic Medline on the YarcData Graph Computer “Urika” and this has been accomplished.

Our second metric is to enable GMU students and professors to use the Semantic Medline – YarcData platform in their work.

This partnerships goal fulfills the need for rapid responses to queries across the entire body of billions of triples or connected data points starting with limited reference locality and in very complex combinations.

Our third metric will be to scale this project to include additional government agencies with relevant Big Data and universities with appropriate data science programs.

2. The multiple stakeholders that will participate in the project and their respective roles and responsibilities;

Semantic Community: Overall coordination, collaboration, and documentation.

National Library of Medicine: Semantic Medline expertise, database, and case studies

YarcData: Graph computer hardware infrastructure and domain expertise including importing RDF triple stores, developing SPARQL queries, and creating visualization tools and dashboards.
GMU: Data science department with students and professors working with large graph databases on government big data sources.

There are at least 2 students who are ready to go on this MEDLine project (i.e., they want to devote most or all of their dissertation research on this, but they will need tutoring by YarcData on the system, instructions on what data you have available, and guidance on where there might be an open research question that they can work on). A YarcData meeting is being set up with these students. A few more students could work on YarcData MEDLine also, specifically for class "big data" assignments (i.e, "not" for research), for this semester.

3. Initial financial and in-kind resources that the stakeholders are prepared to commit to this project; and

All the efforts to date have been donated by the participants.

4. A principal point of contact for the partnership.

Dr. Brand Niemann, Director and Senior Data Scientist, Semantic Community, bniemann@cox.net

---

Previous Presentations

January 24, 2013 Wiki, Wiki, Slides, & Slides

September 10, 2013: Wiki Slides

Tim White: Slides Data Analytics: Supercomputer & Graph Analysis

APRIL 4, 2013: YarcData Announces Winners of $100,000 Graph Analytics Challenge

Tim: Using their own computer for the demo.

YarcData is a subsidiary of Cray because it is a community service business. Graph analytics appliance for government customer who wanted to work with the big data graph.

You don't know what you don't know.

YarcData application for baseball (my Moneyball application) Exponential growth in baseball data. High-risk decisions on $100,000,000 players are made without the benefit of this.

Needle in a haystack: Use a magnet

Stack of Needles: Need a needle piece finder

Why is there a Phillies fan at a Padres game and who is he with?

"I do not know what I am going to search" No reference locality and no starting point.

Federated queries with RDF for discovery - graph analytics is everywhere

DIscovering New Cyber Threats, New Treatments (30 databases around 200 GB - does not work on RDBMS), etc.
Graph Analytics is reaching an inflection Point (Gartner BI Summit, May 2013).

Facebook using Graph Search

Provide a data platform for big data graphs using RDF/SPARQ from W3C.

Real-time, interactive analytics on large graph problems.

Future: Allow for increasingly complex applications and simpler searches.

Tom Rindflesch: Slides: Semantic Medline

What is Semantic Medline and how was it created

Cooperative reciprocity

Exploiting semantic processing by Guiding Research and see Trends

Use Case 1: Schizophrenia

Looking at Glutamate

Use Case 2: Cancer

Looking at Cancer Immunotherapy

Acknowledge Team

Brand Niemann:

We are ready to come back to present our progress to the Federal Big Data Senior Steering Work Group with the story as follows:

1. We formed the Data Science Team and presented what we planned to do at our 14th SOA for eGov Conference, October 2, 2012

http://semanticommunity.info/Federal_SOA/14th_SOA_for_E-Government_Conference_October_2_2012

http://semanticommunity.info/A_NITRD_Dashboard/Semantic_Medline

2. We also presented to the IAC Emerging Technology SIG Meeting: Big Data Committee, November 27, 2012:

http://semanticommunity.info/Emerging_Technology_SIG_Big_Data_Committee

3. We presented our progress to the Federal Big Data Steering Work Group, January 15, 2013 http://semanticommunity.info/Emerging_Technology_SIG_Big_Data_Committee/Government_Challenges_With_Big_Data

https://semanticommunity.info/A_NITRD_Dashboard/Making_the_Most_of_Big_Data

Updated: Thu, 27 Jun 2019 05:55:32 GMT
Powered by MindTouch™
4. We (George Strawn, Tom Rindflesch, and Brand Niemann) discussed how to move forward at lunch: Tom mental health and cancer uses cases, Brand help from YarcData getting the Semantic Medline database running more fully, and George Strawn: more encouragement.

5. YarcData got the Semantic Medline database into their Graph Computer in short order and worked with Tom to explore the two new uses cases that were presented live yesterday to a very enthusiastic and excited audience, September 10, 2013.

http://semanticommunity.info/Data_Science/Cloud_SOA_Semantics_and_Data_Science_Conference (Slides to be posted today) The YarcData Graph Computer was in Wisconsin and the response time to new queries was very fast!

6. We are planning to continue the discussion at the Graph Connect Conference, October 3-4, San Francisco, California, and the data science team work in the Graph Database Meetup, October 22, 2013 (use of Semantic Medline in multiple graph database tools) http://www.graphconnect.com/san-francisco/agenda-san-francisco/

http://www.meetup.com/graphdb-baltimore/events/125172912/

http://semanticommunity.info/Data_Science/Graph_Databases

http://semanticommunity.info/Data_Science/Graph_Databases/Tutorial

---

**Semantic Medline Database Details**


In this page, we provide detailed information about the SemMedDB schema. Database tables, their fields as well as the relationships between the tables are explained. Examples for each table are provided.

**Tables:**

**Name: CITATIONS table**

This table contains relevant metadata for each PubMed citation and has the following data fields:

- **PMID**: PubMed identifier of the citation
- **ISSN**: ISSN identifier of the journal or the proceedings where the article was published
- **DA**: Creation date for the citation
- **DCOM**: Completion date for the citation
- **DP**: Publication date for the citation

<table>
<thead>
<tr>
<th>PMID</th>
<th>ISSN</th>
<th>DA</th>
<th>DCOM</th>
<th>DP</th>
</tr>
</thead>
<tbody>
<tr>
<td>19851774</td>
<td>1432-203X</td>
<td>2010 01 21</td>
<td>2010 03 18</td>
<td>2009 Dec</td>
</tr>
</tbody>
</table>

**Name: CONCEPT table**

This table contains information about the UMLS Metathesaurus concepts as well as EntrezGene terms used by
SemRep. In the current version, UMLS Metathesaurus concepts are from the UMLS 2006AA release. Data fields in this table are as follows:

- **CONCEPT_ID**: Auto generated primary key for each concept
- **CUI**: Concept identifier (CUI) of the concept, corresponds to UMLS CUI if it is from UMLS, and the gene identifier from EntrezGene if it is from EntrezGene
- **TYPE**: "META" if it is a UMLS Metathesaurus concept, "ENTREZ" if it is an EntrezGene symbol
- **PREFERRED_NAME**: UMLS Metathesaurus preferred name for the concept, or the official gene name from EntrezGene
- **GHR**: Corresponding Genetics Home Reference (GHR) identifier, if the concept is a gene or a disorder
- **OMIM**: Corresponding Online Mendelian Inheritance in Men (OMIM) identifier, if the concept is a gene or a disorder

<table>
<thead>
<tr>
<th>CONCEPT_ID</th>
<th>CUI</th>
<th>TYPE</th>
<th>PREFERRED_NAME</th>
<th>GHR</th>
<th>OMIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1844</td>
<td>C0003873</td>
<td>META</td>
<td>Rheumatoid Arthritis</td>
<td>NULL</td>
<td>180300:604302</td>
</tr>
<tr>
<td>1276072</td>
<td>215</td>
<td>ENTREZ</td>
<td>ABCD1</td>
<td>NULL</td>
<td>NULL</td>
</tr>
</tbody>
</table>

Name: **CONCEPT_SEMTYPE** table

This table links concepts in the CONCEPT table with their semantic types. A concept may have multiple semantic types. There is a 1-to-many relation between the CONCEPT and CONCEPT_SEMTYPE tables. The data fields are as follows:

- **CONCEPT_SEMTYPE_ID**: Auto-generated primary key for each concept-semantic type pair
- **CONCEPT_ID**: Foreign key to the CONCEPT table
- **SEMTYPE**: UMLS semantic type abbreviation, such as aapp (Amino Acid, Protein, or Peptide) or gngm (Gene or Genome). For the list of all abbreviations, see [SRDEF](https://semanticommunity.info/A_NITRD_Dashboard/Making_the_Most_of_Big_Data).
- **NOVEL**: Identifies whether the concept is novel or not. Novelty of a concept-semantic type pair is computed based on its distance from root of the UMLS Metathesaurus hierarchy and has been used in automatic summarization approaches based on SemRep [1].

<table>
<thead>
<tr>
<th>CONCEPT_SEMTYPE_ID</th>
<th>CONCEPT_ID</th>
<th>SEMTYPE</th>
<th>NOVEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>2628</td>
<td>1844</td>
<td>dsyn</td>
<td>Y</td>
</tr>
<tr>
<td>1481123</td>
<td>1276072</td>
<td>gngm</td>
<td>Y</td>
</tr>
</tbody>
</table>

Name: **PREDICATION** table

Each record in this table identifies a unique predication. The data fields are as follows:

- **PREDICATION_ID**: Auto-generated primary key for each unique predication
- **PREDICATE**: The string representation of each predicate (for example TREATS, PROCESS_OF)
- **TYPE**: Can be ignored
Name: **PREDICATION_ARGUMENT** table
Each record in this table links a unique predication with one of its arguments. There is a 1-to-many relation between the PREDICATION and PREDICATION_ARGUMENT tables. The data fields are as follows:

- **PREDICATION_ARGUMENT_ID**: Auto-generated primary key for each predication argument
- **PREDICATION_ID**: Foreign key to the PREDICATION table
- **CONCEPT_SEMTYPE_ID**: Foreign key to the CONCEPT_SEMTYPE table
- **TYPE**: ‘S’ for subject argument and ‘O’ for object argument

<table>
<thead>
<tr>
<th>PREDICATION_ARGUMENT_ID</th>
<th>PREDICATION_ID</th>
<th>CONCEPT_SEMTYPE_ID</th>
<th>TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>176604</td>
<td>87120</td>
<td>2628</td>
<td>S</td>
</tr>
<tr>
<td>176605</td>
<td>87120</td>
<td>21437</td>
<td>O</td>
</tr>
</tbody>
</table>

Name: **SENTENCE** table
This table contains information about individual sentences from PubMed citations and includes the following data fields:

- **SENTENCE_ID**: Auto-generated primary key for each sentence
- **PMID**: The PubMed identifier of the citation that the sentence belongs to
- **TYPE**: ‘ti’ for the title of citation and ‘ab’ for the abstract
- **NUMBER**: The location of the sentence within the title or the abstract
- **SENTENCE**: The actual string of this sentence

<table>
<thead>
<tr>
<th>SENTENCE_ID</th>
<th>PMID</th>
<th>TYPE</th>
<th>NUMBER</th>
<th>SENTENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>113049226</td>
<td>19855969</td>
<td>ti</td>
<td>1</td>
<td>Rheumatoid arthritis in patient with homozygous haemoglobin C disease.</td>
</tr>
</tbody>
</table>

Name: **SENTENCE_PREDICATION** table
This table links a sentence with the predications extracted from it. There is a 1-to-many relation between the SENTENCE and SENTENCE_PREDICATION tables. It includes the following data fields:

- **SENTENCE_PREDICATION_ID**: Auto-generated primary key for each sentence-predication pair
- **SENTENCE_ID**: Foreign key to the SENTENCE table
- **PREDICATION_ID**: Foreign key to the PREDICATION table
- **PREDICATION_NUMBER**: The number of times the predication is extracted from the sentence. If there are two instances of the same unique predication in a sentence, the value is 2.
- **CURR_TIMESTAMP**: The timestamp for the record

The rest of the fields in SENTENCE_PREDICATION table provide mention-level information for the elements of the predication (predicate, subject, and object).
- **INDICATOR_TYPE**: The type of the predicate, such as VERB for verbal predicates, and NOM for nominalizations and other argument-taking nouns. For a full list of indicator types, see the Appendix in [2]
- **PREDICATE_START_INDEX**: The first character position of the predicate mention
- **PREDICATE_END_INDEX**: The last character position of the predicate mention
- **SUBJECT_TEXT**: The subject mention in the sentence
- **SUBJECT_DIST**: The distance of the subject mention (counted in noun phrases) from the predicate mention (0 for certain indicator types, such as NOM)
- **SUBJECT_MAXDIST**: The number of potential arguments (in noun phrases) from the predicate mention in the direction of the subject mention (0 for certain indicator types, such as NOM)
- **SUBJECT_START_INDEX**: First character position of the subject mention in the sentence
- **SUBJECT_END_INDEX**: Last character position of the subject mention in the sentence
- **SUBJECT_SCORE**: The confidence score of the mapping between the subject mention and the subject concept
- **OBJECT_***: The fields representing information about the object, in the same way the SUBJECT_* fields do for the subject

<table>
<thead>
<tr>
<th>SENTENCE_PREDICATION_ID</th>
<th>SENTENCE_ID</th>
<th>PREDICATION_ID</th>
<th>PREDICATION_NUMBER</th>
<th>CURR_TIMESTAMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>57109318</td>
<td>113049226</td>
<td>87120</td>
<td>1</td>
<td>2011-11-17 19:58:38.0</td>
</tr>
</tbody>
</table>

Name: **PREDICATION_AGGREGATE** table

This table is a convenience table that joins the salient information from all the above tables for efficient access. The SQL statement that is used to populate this table is as follows:

```sql
insert ignore into PREDICATION_AGGREGATE (PID,SID,PNUMBER, PMID,predicate,s_cui,s_name,s_type,s_novel,o_cui,o_name,o_type,o_novel) select P.PREDICATION_ID, S.SENTENCE_ID, SP.PREDICATION_NUMBER, S.PMID, P.PREDICATE, group_concat(SC.CUI separator '|||'), group_concat(SC.PREFERRED_NAME separator '|||'), group_concat(SS.SEMTYPE separator '|||'), max(SS.NOVEL)=\'Y\', group_concat(OC.CUI separator '|||'), group_concat(OC.PREFERRED_NAME separator '|||'), group_concat(OS.SEMTYPE separator '|||'), max(OS.NOVEL)=\'Y\' from (select * from SENTENCE) S, SENTENCE_PREDICATION SP, PREDICATION P, CONCEPT SC, CONCEPT_SEMTYPE SS, PREDICATION_ARGUMENT SPA, CONCEPT OC, CONCEPT_SEMTYPE OS, PREDICATION_ARGUMENT OPA WHERE S.SENTENCE_ID = SP.SENTENCE_ID and SP.PREDICATION_ID = P.PREDICATION_ID and SPA.PREDICATION_ID = P.PREDICATION_ID and SPA.TYPE=\'S\' AND SS.CONCEPT_SEMTYPE_ID = SPA.CONCEPT_SEMTYPE_ID AND SC.CONCEPT_ID = SS.CONCEPT_ID and OPA.PREDICATION_ID = P.PREDICATION_ID and OPA.TYPE=\'O\' AND OS.CONCEPT_SEMTYPE_ID = OPA.CONCEPT_SEMTYPE_ID AND OC.CONCEPT_ID = OS.CONCEPT_ID group by SP.PREDICATION_ID, SP.SENTENCE_ID, SP.PREDICATION_NUMBER
```

The entity-relationship diagram of SemMedDB is shown below graphically:


---

Google Alters Search to Handle More Complex Queries


By CLAIRE CAIN MILLER

Stephen Lam/Reuters

Amit Singhal, a Google vice president, introduced a search algorithm called Hummingbird on Thursday in Menlo Park, Calif.

Google on Thursday announced one of the biggest changes to its search engine, a rewriting of its algorithm to handle more complex queries that affects 90 percent of all searches.

The change, which represents a new approach to search for Google, required the biggest changes to the company’s search algorithm since 2000. Now, Google, the world’s most popular search engine, will focus more on trying to understand the meanings of and relationships among things, as opposed to its original strategy of matching keywords.
The company made the changes, executives said, because Google users are asking increasingly long and complex questions and are searching Google more often on mobile phones with voice search.

“They said, ‘Let’s go back and basically replace the engine of a 1950s car,’ ” said Danny Sullivan, founding editor of Search Engine Land, an industry blog. “It’s fair to say the general public seemed not to have noticed that Google ripped out its engine while driving down the road and replaced it with something else.”

Google announced the new algorithm, called Hummingbird, at an event to celebrate the search engine’s 15th birthday. The event was held in the garage Google’s founders rented when they started the company. Google revealed few details about how the new algorithm works or what it changed. It said it made the change a month ago, though consumers may not have noticed a significant difference to search results during that time.

Google originally matched keywords in a search query to the same words on Web pages. Hummingbird is the culmination of a shift to understanding the meaning of phrases in a query and displaying Web pages that more accurately match that meaning.

Google had taken smaller steps toward this. The Knowledge Graph, introduced last year, understands the meanings of and relationships between things, people and places, which is known as semantic search. It is why a search for Michelle Obama, for instance, shows her birthday, hometown and family members’ names, as well as links to related people like Hillary Rodham Clinton and Joseph R. Biden Jr.

The algorithm also builds on work Google has done to understand conversational language, like interpreting what pronouns in a search query refer to.

Hummingbird extends that to all Web searches, not just results related to entities included in the Knowledge Graph. It tries to connect phrases and understand concepts in a long query.

The outcome is not a change in how Google searches the Web, but in the results that it shows. Unlike some of its other algorithm changes, including one that pushed down so-called content farms in search results, Hummingbird is unlikely to noticeably affect certain categories of Web businesses, Mr. Sullivan said. Instead, Google says it believes that users will see more precise results.

Google also announced a few smaller changes to searching. It is changing the visual layout of mobile search to better suit phones and tablets. People can now compare two things, like butter and olive oil, or corgis and pugs, in search results. And with a new app for Apple devices, people can set reminders on an Android device at home and receive them later on an iPhone.

A version of this article appears in print on 09/27/2013, on page B2 of the New York edition with the headline: Google Unveils a New Approach to Searches.

My Comment

So Google says: We are in the early phases of moving from an information engine to a knowledge engine in the video: http://www.google.com/insidesearch/features/search/knowledge.html
entitled: Learn more about the Knowledge Graph. Go under the hood with members of the team who work on this technology.

Google says: Their new focus is more on trying to understand the meanings of and relationships among things, as opposed to its original strategy of matching keywords, because users are asking increasingly long and complex questions and are searching Google more often on mobile phones with voice search.

I say: This is because of Apples' Siri which we know has had at least 10 years of serious development using real semantics by Tom Gruber:

http://en.wikipedia.org/wiki/Tom_Gruber

Google says: It tries to connect phrases and understand concepts in a long query. The outcome is not a change in how Google searches the Web, but in the results that it shows. Users will see more precise results. We can use the Knowledge Graph to answer questions you never thought to ask and help you discover more.

I say: This is what Jim Hendler says: A little semantics goes a long way. But not as far as what Tom has done with Semantic Medline and now running on the YarcData Appliance.

YarcData says: On the Watson question (how does YarcData compare to Watson), Sandia did tests comparing our hardware, XMT with 4 processors against Blue Gene/L with 32,000 processors. On a large graph we ran shortest path in half the time. The smallest machine we build now is 64 processors.

I say: Knowledge Base Semantic Search can be demonstrated on a small, but practical scale, the way I do in my data science work.

You put the related content with semantic links into a MindTouch page like: http://semanticommunity.info/Data_Science/Data_Science_Symposium_2013

You use Google Chrome Find for the word or phrase you are looking for.

You build an index of triples in a spreadsheet of that MindTouch page and combine it with similar indices of other MindTouch pages and relational tables, all of which are then searchable in Excel and in Spotfire where you can do analytics on them and allow people to search them on the Web as well.

Spotfire contains a module called Network Analytics developed for the CIA that not only displays the network graphs like Tom does in Semantic Medline, but also calculates their statistics, like Facebook has done recently:


So it is all about structuring content so it is data with strong relationships, and if the format is triples and you have a scalable way to display the search of those triples, then you have Knowledge Base Semantic Search on a small (me) or large scale (Google, etc.). Even better is if you have your content relationships in a native, scalable, graph database, you get more precise (succinct and accurate) results which is the goal of all of this as Google says: to answer questions you never thought to ask and help you discover more.

Tom demonstrated this on September 10th and we are trying to get Semantic Medline into Neo4j (the leading native, scalable graph database), and that is why we are excited!
Aiming to make the most of the explosion of Big Data and the tools needed to analyze it, the Obama Administration announced a “National Big Data Research and Development Initiative” on March 29, 2012. To launch the initiative, six Federal departments and agencies announced more than $200 million in new commitments that, together, promise to greatly improve and develop the tools, techniques, and human capital needed to move from data to knowledge to action. The Administration is also working to “liberate” government data and voluntarily-contributed corporate data to fuel entrepreneurship, create jobs, and improve the lives of Americans in tangible ways. For additional information about the launch of the Big Data Initiative see the OSTP Fact Sheet and Press Release.

As we enter the second year of the Big Data Initiative, the Administration is encouraging multiple stakeholders including federal agencies, private industry, academia, state and local government, non-profits, and foundations, to develop and participate in Big Data innovation projects across the country. Later this year, the Office of Science and Technology Policy (OSTP), NSF, and other agencies in the Networking and Information Technology R&D (NITRD) program (My Note: See below) plan to convene an event that highlights high-impact collaborations and identifies areas for expanded collaboration between the public and private sectors. The Administration is particularly interested in projects and initiatives that:

- Advance technologies that support Big Data and data analytics;
- Educate and expand the Big Data workforce;
- Develop, demonstrate and evaluate applications of Big Data that improve key outcomes in economic growth, job creation, education, health, energy, sustainability, public safety, advanced manufacturing, science and engineering, and global development;
- Demonstrate the role that prizes and challenges can play in deriving new insights from Big Data; and
- Foster regional innovation.

Please submit a two-page summary of projects to BigDataprojects@nitrd.gov. The summary should identify:

1. The goal of the project, with metrics for evaluating the success or failure of the project;
2. The multiple stakeholders that will participate in the project and their respective roles and responsibilities;
3. Initial financial and in-kind resources that the stakeholders are prepared to commit to this project; and
4. A principal point of contact for the partnership.

The submission should also indicate whether NITRD can post the project description to a public website. Unless otherwise noted, submissions with sensitive material (e.g., trade secrets, or privileged or confidential commercial or financial information) will be protected from disclosure.

This announcement is posted solely for information and planning purposes; it does not constitute a formal solicitation for grants, contracts, or cooperative agreements.

Deadline Date for Submission of Summaries: September 2, 2013
The National Coordination Office (NCO) for Networking and Information Technology Research and Development (NITRD) supports the planning, budget, and assessment activities for the Federal government's NITRD Program.

The NITRD Program, chartered by Federal law, is the primary mechanism by which the Government coordinates its unclassified networking and information technology (IT) research and development (R&D) investments. Eighteen Federal agencies, including all of the large science and technology agencies, are formal members of the NITRD Program. Many other Federal organizations also participate in NITRD activities.

These agencies work together to develop a broad spectrum of advanced networking and IT capabilities to power Federal missions; U.S. science, engineering, and technology leadership; and U.S. economic competitiveness. Their efforts increase the overall effectiveness and productivity of Federal networking and IT R&D investments, leveraging strengths, avoiding duplication, and increasing interoperability of networking and IT R&D products.

Key NCO activities include the following:

- The NCO supports the NITRD Subcommittee, which coordinates the NITRD Program, and the organizations that report to the Subcommittee. The Subcommittee reports through the Committee on Technology to the Cabinet-level National Science and Technology Council. The organizations that report to the Subcommittee include Interagency Working Groups (IWGs) and Coordinating Groups (CGs) in the following R&D areas:
  - BD - Big Data (My Note: See below)
  - CSIA - Cyber Security and Information Assurance
  - HCI&IM - Human Computer Interaction and Information Management
  - HCSS - High Confidence Software and Systems
  - HEC - High End Computing
  - HITRD - Health Information Technology Research and Development
  - LSN - Large Scale Networking
  - SDP - Software Design and Productivity
  - SEW - Social, Economic, and Workforce Implications of IT and IT Workforce Development
  - WSRD - Wireless Spectrum Research and Development

- In cooperation with the NITRD agencies, IWGs, and CGs, the NCO prepares, publishes, and disseminates the Program's annual supplement to the President's Budget, Federal networking and IT R&D plans, and networking and IT research needs reports.

- The NCO works closely with the Office of Science and Technology Policy (OSTP) and the Office of Management and Budget in the Executive Office of the President on NITRD policy and budget issues. The Director of the NCO is appointed by the Director of OSTP.

- The NCO provides technical support for the activities of the Networking and Information Technology Subcommittee of the President's Council of Advisors on Science and Technology, a panel of experts from industry and academia, in assessing the NITRD Program and preparing associated reports.

Big Data

The Big Data Senior Steering Group (BD SSG) works to facilitate and further the goals of the White House Big Data R&D Initiative.

Big Data Initiative

Aiming to make the most of the fast-growing volume of digital data, in March 2012, the Obama Administration announced the "Big Data Research and Development Initiative." By improving our ability to extract knowledge and insights from large and complex collections of digital data, the initiative promises to help solve some the Nation’s most pressing challenges. The Big Data Senior Steering Group continues to work with the White House Office of Science and Technology along with several Federal departments and agencies to help further the goals of the Initiative which strive to:

- Advance state-of-the-art core technologies needed to collect, store, preserve, manage, analyze, and share huge quantities of data.
- Harness these technologies to accelerate the pace of discovery in science and engineering, strengthen our national security, and transform teaching and learning; and
- Expand the workforce needed to develop and use Big Data technologies.

For additional information, see the White House Press Release (PDF), and the Big Data Fact Sheet(PDF). (My Note: I have these elsewhere)

http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf

http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_fact_sheet_final.pdf

Call for Projects

- Click here to read the Call for Projects (My Note: See above)
Big Data Events and Workshops

**May 29, 2013 - Data Sharing and Metadata Curation: Obstacles and Strategies:** Future strategies for managing scientific data and metadata for basic and applied research

[Workshop Summary Report](#) (My Note: See below)

The purpose of this workshop was to have focused discussions on future strategies for managing scientific data and metadata for basic and applied research; specifically,

(a) how to better enable, encourage, and realize sharing of data, both across disciplinary divides and between “micro-silos” within research domains,

(b) how to acquire, manage, and curate metadata in order to ensure usability and comprehensibility of data over time and between disciplines, and

(c) how to enable data discovery, access, and analysis across distributed, public, and private data centers.

**May 3, 2013 - White House Big Data Partners Workshop** (My Note: See below)

Location: White House Conference Center

As we enter the second year of the National Big Data Research and Development Initiative (launched on March 29, 2012), the Obama Administration is encouraging multiple stakeholders—including Federal agencies, private industry, academia, state and local government, non-profits, and foundations—to develop and participate in Big Data innovation projects across the country. To surface such partnerships and novel approaches that would allow partnerships to scale across the country, a series of workshops and events are being planned over the spring and summer 2013.

On May 3, 2013, the first workshop brought together representatives from industry, academia, and government to learn about existing BD partnerships, make connections with interested parties, and explore future possibilities.

**Overview**

*Big Data is a term applied to data sets whose size is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time. Big data sizes are a constantly moving target currently ranging from a few dozen terabytes to many petabytes of data in a single data set.* – Wikipedia, May 2011

The Big Data Senior Steering Group (BDSSG) was formed to identify current big data research and development activities across the Federal government, offer opportunities for coordination, and identify what the goal of a national initiative in this area would look like. Subsequently, in March 2012, The White House Big Data R&D Initiative was launched and the BDSSG continues to work in four main areas to facilitate and further the goals of the Initiative.

**Four Main Areas and their current activity**

- **Core Technologies** – An interagency Big Data Solicitation is ongoing at: [http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504767](http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504767) (My Note: I have this somewhere else)
• **Domain Research Projects** – Specific research projects that involve the use of big data and interagency collaboration have been identified. The group is currently working to identify common data infrastructure challenges that would benefit from additional research.

• **Challenges and Competitions** – A Big Data Ideation Challenge began in October 2012 as the first in a series of contests that will build understanding about “taking diverse and heterogeneous data sets and making them more homogeneous and usable”. The information derived from this series will inform recommendations on big data research.

• **Workforce Development** – Current Agency workforce development programs that either include, or could be evolved to include, big data training have been identified. Currently, a publicly accessible web portal to disseminate this and other information on Federal resources such as access to computing, networking, and storage capacity, is underway. Information is also being collected on current models for creating public-private partnerships which could lead to new program recommendations.

---

**In the news**

• **March 14 (Bloomberg)** -- Altman, Groschupf, Mahony, Strawn `Big Data' Value

"Anne Altman, a general manager at International Business Machines Corp., Stefan Groschupf, co-founder and chief executive officer of Datameer Inc., Colin Mahony, senior vice president at Vertica, a unit of Hewlett-Packard Co., and George Strawn, a director at the National Coordination Office, participate in a panel discussion about opportunities in so-called big data services. Bloomberg's Michael Nelson moderates the panel at the Bloomberg Link Big Data Conference in Washington. (Source: Bloomberg)"

[watch video](#)

**Presentations**

*January 22, 2013 from 12:30 to 1:30; in NSF Stafford II-555*

Presentation and discussion with Professor Joel Cutcher-Gershenfeld. Professor Cutcher-Gershenfeld will be presenting information on the NSF EarthCube initiative including stakeholder survey data (approximately 850 responses). For more info [click here](#)

**My Note: Our Data Science Team presented on January 24, 2013** [Wiki, Wiki, Slides, & Slides](#)
White House Big Data Partners Workshop


May 3, 2013, White House Conference Center

Purpose Statement

As we enter the second year of the National Big Data Research and Development Initiative (launched on March 29, 2012), the Obama Administration is encouraging multiple stakeholders—including Federal agencies, private industry, academia, state and local government, non-profits, and foundations—to develop and participate in Big Data innovation projects across the country. To surface such partnerships and novel approaches that would allow partnerships to scale across the country, a series of workshops and events are being planned over the spring and summer 2013.

On May 3, 2013, the first workshop brought together representatives from industry, academia, and government to learn about existing BD partnerships, make connections with interested parties, and explore future possibilities. Below you will find presentations from the workshop participants and sponsors.

Workshop Agenda

[White House Big Data Partners Workshop Agenda](#) (PDF)

Workshop Report

document pending

Introductory Speaker Presentations

[Unleashing the Power of Big Data: Why We Need All Hands on Deck](#) (PDF)

Tom Kalil - Deputy Director, Division of Technology and Innovation Office of Science and Technology Policy

[The Promise of Big Data](#) (PDF)

Farnam Jahanian - Co-Chair, Networking & Information Technology Research & Development (NITRD) Subcommittee
Assistant Director for the Computer and Information Science and Engineering (CISE) Directorate National Science Foundation

[Big Data Partners Workshop remarks by Suzi Iacono](#) (PDF)

Suzanne Iacono - Co-Chair, Big Data Senior Steering Group, NITRD National Science Foundation
Panel Speaker Presentations

- **Building a Data-Driven Society** *(PDF)*
  Jake Porway - *Founder and Executive Director DataKind*

- **Data Science & Multi-Stakeholder Partnerships** *(PDF)*
  Annika Jimenez - *Global Head of Data Science Services Pivotal*

Mark Peterson - *Team Lead, Data and Analytics, Office of Science & Technology U.S. Agency for International Development*

**My Note:** No slides

- **RENCI Collaborations and Consortia to Advance Data Science** *(PDF)*
  Stanley Ahalt - *Director, Renaissance Computing Institute University of North Carolina at Chapel Hill*

Lightning Round Presentations

- **Lightning Talk Slides** *(PDF)*

Relevant Documents/Links

- Fact Sheet: Big Data Across the Federal Government (March 29, 2012) *(My Note: See above)*
- Unleashing the Power of Big Data - [http://www.whitehouse.gov/blog/2013/04/18/unleashing-power-big-data](http://www.whitehouse.gov/blog/2013/04/18/unleashing-power-big-data) *(My Note: See below)*
- “BIG DATA” Initiative *(My Note: See above)*

March 29, 2012. View the March 29, 2012 webcast of the Federal Government Big Data Rollout. View video interviews with Farnam Jahanian, assistant director ...

Event BIGDATA Webinar. Core Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA). May 8, 2012 11:00 AM to. May 8 ...

NSF Announces Interagency Progress on Administration's Big Data Initiative *(My Note: This link is broken)*

Oct 3, 2012 - The National Science Foundation (NSF), with support from the National Institutes of Health (NIH), today announced nearly $15 million in new Big Data fundamental research projects...

Unleashing the Power of Big Data

Source: [http://www.whitehouse.gov/blog/2013/04/18/unleashing-power-big-data](http://www.whitehouse.gov/blog/2013/04/18/unleashing-power-big-data)
Posted by Tom Kalil and Fen Zhao on April 18, 2013 at 04:04 PM EDT

Last year, the Obama Administration announced the [National Big Data Research and Development Initiative](http://www.whitehouse.gov/blog/2013/04/18/unleashing-power-big-data)—a major step toward addressing the challenge and opportunity of “Big Data.” Big Data are data sets so large, complex, or rapidly-generated that they can’t be processed by traditional information and communication technologies. At its launch, the Big Data Initiative featured more than $200 million in new commitments from six Federal departments and agencies aiming to make the most of the explosion of Big Data and the tools needed to analyze it.

Every day, decision makers, resource manager, engineers, first-responders, scientists, and citizens are faced with a multitude of constantly flowing data streams coming from many sources, in many formats. Making sense of these volumes of Big Data requires cutting-edge tools and technologies that can analyze and extract useful knowledge from vast and diverse streams of information. Wrapping our arms around Big Data could lead to an array of important societal benefits—from empowering consumers with the full landscape of information they need to make optimal energy decisions; to enabling civil engineers to monitor and identify at-risk infrastructure; to informing more accurate predictions of natural disasters; and more.

As we enter the second year of the Big Data Initiative, the Obama Administration is encouraging multiple stakeholders, including federal agencies, private industry, academia, state and local government, non-profits, and foundations to develop and participate in Big Data initiatives across the country. Of particular interest are partnerships designed to advance core Big Data technologies; harness the power of Big Data to advance national goals such as economic growth, education, health, and clean energy; use competitions and challenges; and foster regional innovation.

The National Science Foundation has issued a [request for information](http://www.whitehouse.gov/blog/2013/04/18/unleashing-power-big-data) encouraging stakeholders to identify Big Data projects they would be willing to support to achieve these goals. And, later this year, OSTP, NSF, and other partner agencies in the [Networking and Information Technology R&D](http://www.whitehouse.gov/blog/2013/04/18/unleashing-power-big-data) (NITRD) program plan to convene an event that highlights high-impact collaborations and identifies areas for expanded collaboration between the public and private sectors.

*Tom Kalil is Deputy Director for Technology and Innovation at OSTP*

*Fen Zhao is an AAAS Fellow at OSTP*

**Workshop Agenda**

[White House Big Data Partners Workshop Agenda](http://www.whitehouse.gov/blog/2013/04/18/unleashing-power-big-data) (PDF)

830-900 Check-in for All Attendees Lincoln Room
PLENARY SESSION Truman Room

900-930 Opening Remarks
Tom Kalil, Deputy Director, Division of Technology and Innovation Office of Science and Technology Policy
Farnam Jahanian Co-Chair, Networking & Information Technology Research & Development (NITRD) Subcommittee
Assistant Director for the Computer and Information Science and Engineering (CISE) Directorate National Science Foundation
Suzanne Iacono Co-Chair, Big Data Senior Steering Group, NITRD National Science Foundation

930-1100 Panel Discussion
Peter Lyster (Moderator) Program Director, Division of Biomedical Technology, Bioinformatics, and Computational Biology National Institutes of Health
Jake Porway Founder and Executive Director DataKind
Annika Jimenez Global Head of Data Science Services Pivotal
Michael Rappa Executive Director, Institute for Advanced Analytics North Carolina State University
Mark Peterson Team Lead, Data and Analytics, Office of Science & Technology U.S. Agency for International Development
Stanley Ahalt Director, Renaissance Computing Institute University of North Carolina at Chapel Hill

1100-1200 Lightning Round Presentations,
1 minute presentations by workshop attendees on projects of interest with Q&A, facilitated by
Fen Zhao AAAS Science & Technology Policy Fellow Office of Science and Technology Policy
Jonathan Petters AAAS Science & Technology Policy Fellow Department of Energy

1200-115 Lunch Break
Directions to nearby eateries are provided.

130-245 Beakout Sessions

Truman Room Partnerships for Innovation, facilitated by
Mark Suskin Deputy Division Director, Division of Advanced Cyberinfrastructure National Science Foundation
Barbara Helland Associate Director, Advanced Scientific Computing Research Department of Energy

Jackson Room Community Engagement, facilitated by
Jason Crusan Director, Advanced Exploration Systems Division National Aeronautics and Space Administration
Jeanne Holm Evangelist, Data.gov U.S. General Services Administration

Lincoln Room Education and Workforce, facilitated by
Michelle Dunn Program Director, National Cancer Institute National Institutes of Health
Nandini Kannan Program Director, Division of Mathematical Sciences National Science Foundation

245-300 Break

PLENARY SESSION RECONVENES Truman Room

300-345 Break-Out Sessions Report Back, facilitated by
Wendy Wigen Coordinator, Big Data Senior Steering Group, NITRD National Coordination Office

345-400 Closing Remarks

George Strawn Director, Federal Networking and Information Technology Research and Development Program National Coordination Office

Todd Park United States Chief Technology Officer Office of Science and Technology Policy
Unleashing the Power of Big Data: Why We Need All Hands on Deck

Tom Kalil - Deputy Director, Division of Technology and Innovation Office of Science and Technology Policy (PDF)

Title Slide

Unleashing the Power of Big Data: Why We Need All Hands on Deck

Thomas Kalil
Deputy Director for Technology and Innovation
White House Office of Science and Technology Policy & National Economic Council
thkalil@ostp.eop.gov
May 3, 2013

Administration interest in Big Data 1

Administration interest in Big Data (1)

• Economic impact – demonstrated productivity increases for firms that are mastering Big Data and analytics

• Accelerate pace of discovery in science and engineering

• National security/intelligence

• Health – better care at lower costs, new infrastructure for clinical research + drug discovery that integrates EHRs, m-health, omics
Administration interest in Big Data 2

Administration interest in Big Data (2)

• Education:
  – Move from data-poor to data-rich domain.
  – Online courses that improve the more students use them.
  – Online courses as “wind tunnel” for learning science

• Smart grid, energy efficiency

• Real-time labor market information

Government role 1

Government role (1)

• Invest in R&D related to Big Data technologies

• Support efforts to expand “Big Data” workforce (e.g. data scientists)

• Support applications of Big Data in national priorities and for agency missions, be a “smart user” of Big Data approaches

• Use challenges to demonstrate what is possible (see, e.g., http://www.kaggle.com and http://www.topcoder.com)
Government role 2

Government role (2)

• Policy issues:
  – Privacy
  – Transborder data flow
  – Global applications (see, e.g. UN Global Pulse)

• Make more government data available in bulk-downloadable, machine-readable format consistent with President’s Open Government Initiative

• Be a catalyst for investments by private sector

March 29, 2012 announcement

March 29, 2012 announcement

• Over $200 million in investments in R&D related to Big Data, both grants and solicitations
• DARPA XDATA program
• Joint solicitation between National Science Foundation and National Institutes of Health
• Department of Energy-funded institute in scientific visualization
• NIH 1,000 Genome Project, joint with European Bioinformatics Institute
All Hands on Deck

- Make the “big data” effort a national initiative as opposed to a federal initiative.
- Identify steps that all stakeholders (companies, regions, early adopters, universities, researchers, investors, non-profits, foundations, professional societies, skilled volunteers) have taken and can take to make the most of the opportunities created by Big Data
- High-profile event in Fall 2013 to celebrate and catalyze new and expanded commitments

Examples of actions to advance Big Data 1

- Universities:
  - New courses and courses of study – e.g. CMU Department of Machine Learning, Northwestern Master’s in Analytics, NCSU MSA
  - MOOCs
- University-industry collaborations
  - MIT + Intel – bigdata@CSAIL
  - UC Berkeley – AMPLAB – NSF + 17 industry collaborators
- Regional efforts
  - Massachusetts Big Data Initiative
- Data challenges (e.g. Hewlett Foundation on automated student assessment)
Examples of actions to advance Big Data 2

Examples of actions to advance Big Data (2)

- Open sourcing Big Data software, publishing papers on key results from private R&D (e.g. Hadoop, R, Hbase, Cassandra, Drill, etc.)

- Data philanthropy
  - Orange making 2.5 billion anonymized cellphone records from Cote D’Ivoire available to researchers

- Collaboration with Big Data oriented non-profits (e.g. Datakind)

Examples of actions to advance Big Data 3

Examples of actions to advance Big Data (3)

- Collaborations to demonstrate value of Big Data (e.g. treat and prevent risk factors for stroke, heart attack, diabetes)
  - Pilots with rigorous evaluation
  - Strategies for scaling if successful

- Help with national agenda-setting (e.g. Computing Community Consortium, TechAmerica Foundation)

- Your idea here

The Promise of Big Data

Farnam Jahanian - Co-Chair, Networking & Information Technology Research & Development (NITRD) Subcommittee
Assistant Director for the Computer and Information Science and Engineering (CISE) Directorate National Science Foundation (PDF)

Big Data Partners Workshop remarks by Suzi Iacono

Suzanne Iacono - Co-Chair, Big Data Senior Steering Group, NITRD National Science Foundation (PDF)
Big Data Partners Workshop
  Suzi Iacono, Co-chair
  Big Data Senior Steering
  Group

Big Data Workshop 1

Big Data Workshop

• Who is here?
• A mix: a multi-stakeholder group
  – Industry
  – Academia
  – Not for profits
  – Associations
  – Government
• Opportunity to talk with others who may be doing things differently than what you are doing, but who also care passionately about Big Data

Big Data Workshop 2
Big Data Workshop

- Why are you here?
- You have made some connection about Big Data with OSTP, the Big Data Senior Steering Group and/or one of the agencies involved
- The plan: Bring everyone together for brainstorming, discussion, planning
- Learn about multi-stakeholder partnerships
- Begin to think about how you can make a difference

Big Data Workshop 3

Big Data Workshop

- What will you do?
- Morning – listen to: different points of view, project results, new ideas, new partnerships, how others are innovating, etc.
- Afternoon – engage in information exchange: brainstorm, explore possibilities and come up with new ideas for what YOU will do, who you will talk to here and once you leave
Multi-stakeholder Partnerships

- Talk to those that you would not normally talk to
- Establish connections with other stakeholder groups – within a region, across a broad sector, etc.
- Discern how you could make significant positive change with Big Data
- Take risks that you would not normally take

Charge

- Fact finding: Collect data and information
- Idea finding: Listen for new ideas, models, partnerships, etc.
- Partner finding: Search for your Big Data “soulmates” (or partners like Gilbert and Sullivan, Fred and Ginger)
- Solution finding: Discern promising ideas that can be applied and that would make a difference
Avoid these Idea Killers

Avoid these Idea Killers

- We tried that already
- We’ve never done that before
- We don’t do it that way here
- Not in our budget
- Not an interesting problem
- We don’t have time
- Management will never go for it
- It’s out of scope
- People won’t like it
- It won’t make enough money
- How stupid are you?
- You are smarter with your mouth shut

Logistics

Logistics

- Coffee and water – Thanks to Sage BioNetworks
- Lunch – On your own, after the lightening talks, please go in groups; there are handouts to find restaurants
- Video cameos – Huge response, some today, some at later workshops
- Breakouts – After lunch, go to one of three sessions –
  - Partnerships for Innovation – Truman Room
  - Community Engagement – Jackson Room
  - Education and Workforce – Lincoln Room
- Big Data web pages – videos, report from this workshop, previous press releases, OSTP Fact Sheet, etc.
- Huge thanks – OSTP, NCO, Big Data SSG, and everyone else who lent a helping hand!
Parting Words

- Think big, be bold.
- “...ordinary things, people, and events are transformed into legends by the forces of time” — Berkun, Scott. The Myths of Innovation, Sebastopol: O'Reilly Media, Inc., 2010.
- “Innovators are the test pilots of life, taking big chances so we don’t have to” — Berkun, Scott. The Myths of Innovation, Sebastopol: O'Reilly Media, Inc., 2010.

Panel Speaker Presentations

**Building a Data-Driven Society**

Jake Porway - Founder and Executive Director DataKind ([PDF](https://semanticommunity.info/A_NITRD_Dashboard/Making_the_Most_of_Big_Data))

**Data Science & Multi-Stakeholder Partnerships**

Annika Jimenez - Global Head of Data Science Services Pivotal ([PDF](https://semanticommunity.info/A_NITRD_Dashboard/Making_the_Most_of_Big_Data))

Mark Peterson - Team Lead, Data and Analytics, Office of Science & Technology U.S. Agency for International Development

**RENCI Collaborations and Consortia to Advance Data Science**

Stanley Ahalt - Director, Renaissance Computing Institute University of North Carolina at Chapel Hill ([PDF](https://semanticommunity.info/A_NITRD_Dashboard/Making_the_Most_of_Big_Data))

Lightning Round Presentations

Lightning Talk Slides ([PDF](https://semanticommunity.info/A_NITRD_Dashboard/Making_the_Most_of_Big_Data))
Genomes Galore – Big Data Analytics for High Throughput DNA Sequencing

Drive Grand Challenges
- Identification of complex disease traits
- Detection of biological threats
- Microbial studies and human health
- Plant genotype to phenotype

Vision and Goals
- Empower community migration to HPC
- Preserve ability to create new solutions
- Target researchers & software developers

Research and Dissemination Approach

Write Optimization: The Future of Big Data Storage

Write optimization impact on SSDs:
- Better wear-out.
- Better compression.

MongDB vs Fractal Tree index

NSF Big Data grant: Algorithmics of write-optimization.
NSF & DOE: Supported tech transfer.

Write-optimized Fractal Tree indexes ingest data 10x-100x faster.

Index maintenance has been notoriously slow.

Write optimization impact on query I/O load

Updated: Thu, 27 Jun 2019 05:55:32 GMT
Powered by mindtouch™
MIT Big Data Initiative at CSAIL

**Mission**: The goal of the MIT Big Data Initiative at CSAIL is to collaborate with industry to develop new technologies and tools to solve the next generation data challenges. We want to enable individuals and organizations to truly leverage Big Data to develop new tools and platforms that are used in multiple application domains.

**Industry Members**

- AIG
- EMC
- Facebook
- BT
- Nokia
- SAP
- Microsoft
- Huawei
- Intel

**http://bigdata.csail.mit.edu/**

Big Data Science for the Masses: start small, Think Big

**Visualize This:**

A sea of Data (sea of CDs)

This is the CD Sea in Kimlington, England (600,000 CDs = 300 TB)

“Big Data” is different!

We need more Data Scientists in order to discover the unknown unknowns in BIG DATA collections more efficiently and more effectively.

1) **Informatics in Education**

Work with data in all learning settings:
- Informatics (Data Science) enables transparent reuse and analysis of data in inquiry-based classroom learning.
- Learning is enhanced when students work with real data and information (especially online data) that are related to the topic (any topic) being studied.
- http://www.cscatskills.org/usingdata/ (“Using Data in the Classroom”)

2) **An Education in Informatics**

Students are specifically trained to:
- access & query large distributed data repositories;
- conduct meaningful inquiries into data;
- mine, visualize, and analyze the data;
- make objective data-driven inferences, discoveries, and decisions; and
- communicate “stories” through data.
Self-Generated Health Information Exchange

Open Science for Big Data
Dataverse for Big Data

The Dataverse for Big Data will make very large data sets sharable, citable, and searchable, and facilitate reproducibility of research emerging from this new generation of data. In some cases, the largest data sets are now being used by researchers, and others are working on new applications that will require even larger data sets. The Dataverse provides a solution to find, share, cite, and reuse them, having the largest repository of social science data in the world and replicated across institutions worldwide. However, a small but increasing number of very large data sets are now being used by researchers, offering a rich new source of data and new insights about the world. Now, data sets may exceed the amount of data that can be transferred to and stored on the researcher's computer. Another challenge is that the amount of data cannot be any longer controlled and analyzed by the file that collected them, and therefore a substantial fraction of the data remains unexplored. And a third challenge is that other data sets contain privacy sensitive data. The Dataverse for Big Data will meet these challenges by providing an open science platform where all researchers will be able to explore, analyze and query (real-time, in some cases) big data sets, predictively cite a subset of the data, get the provenance trail of a data set, use privacy and secure tools to explore sensitive data, and when needed, use crowdsourcing tools to bring additional people to help curate and analyze the data.

NOW
Large # of small data sets
Disconnected, fixed-size data
Users access multiple formats
Downloadable, fixed-size data
Faxes, paper

COMING SOON
Small # of very large data sets
On the fly, efficiently
Provider’s data
Privacy Preserving Real-time Data
Crowdsourcing

Big Data (TB – EB)

Frequently Updated
Data Streaming

Privacy Sensitive

The Dataverse platform will integrate with other technologies: Foreman Data – a system to harvest frequently updated data sets... and MyDataCan – a system to securely store personal data that eventually can be donated to science. The open science platform will enable distribution of Big Data from a variety of scientific fields. The Dataverse platform will also enable data integration and sharing across fields, such as astronomy observations from nearby approaching objects from the Pan-STARRS project, health data from hospital clinics, student usage data collected by measures online open courses such as EdX, and social media data.

1 Primary contact: Venkataramanan, venkat@lirneasia.org, datav.org

University of Kentucky

SAP & Dell help promote high graduation

Dell is helping UK use SAP's HANA to quickly and accurately identify students at risk of leaving the institution. Earlier intervention will help keep UK's students on track.

HANA will provide UK with the ability to gain new insight into its student body and could allow better targeting of at-risk students. This effort benefits both UK's students and Kentucky's taxpayers, for every 1% increase in UK's graduation rate, the University potentially gains over $1 million through tuition and increased earnings capacity.

https://semanticommunity.info/A_NITRD_Dashboard/Making_the_Most_of_Big_Data
Updated: Thu, 27 Jun 2019 05:55:32 GMT
Powered by MindTouch™
Synergistic Co-Design for BIG DATA

Co-Design Exemplars
1. PanMEDIC: Parallel Metadata Environment for Distributed I/O & Computing (ps → ms) → Find missing genes
   http://archives.nrg.org/lsdb/100811
2. Molecular Modeling → Rational drug design
   http://www.molecular.com/washt/47fBFewVq5h
3. Temporal Data Mining of Brain
   http://www.nerc.edu/usa/hs/online-temporal-dimension.html
   geno-2011.pdf

Data Science Incubator

Incubator
- Seed grants to students and postdocs
- Rotating staff from science and industry
- An evolving portfolio of reusable services
- A network of cross-boundary partnerships
- Produce digital capital and human capital

Some local observations:
- Big data work exposes common ground
- Every job is becoming “data scientist”
- More IT-shaped people!
- Democratization to the long tail is key
- Industry and research aren’t too different

https://semanticommunity.info/A_NITRD_Dashboard/Making_the_Most_of_Big_Data
Updated: Thu, 27 Jun 2019 05:55:32 GMT
Powered by mindtouch
MITRE's Big Data Analytics Activities

- The MITRE Corporation is a not for profit company that runs multiple FFRCs, chartered in the public interest
- We have expertise in the full range of big data technologies, including high performance computing, complex event processing, parallel relational databases and analytic cloud computing
- The MITRE Innovation Program invests in game-changing research: cyber security, healthcare informatics, counter fraud, enhancing intelligence analysis, and core computational approaches for big data
  - Our researchers partner with government, academia, industry
- The MITRE Institute runs a series of internal training courses on big data and cloud computing
- MITRE's Enterprise Computing Environment provide on-demand computing and storage services to our projects
- We work across missions and agencies


Globus Online

It should be trivial for all researchers to Collect, Organize, Move, Sync, Share, Analyze, Annotate, Publish, Search, Backup, & Archive BIG DATA ...but in reality it's very challenging

Globus Online uses SaaS approaches to address this challenge and make advanced research data management capabilities broadly accessible

Big Data management + NGS Analysis pipeline + Cloud Computing Infrastructure = Flexible, scalable, easy-to-use genomics analysis for all biologists
Confidentiality and Data Access in the Use of Big Data: Theory and Practical Approaches

Confidentiality and Data Access in the Use of Big Data: Theory and Practical Approaches
Editors: Stefan Bender, Julie Lane, Helene Rosenbaum, Victoria Stodden

Goals
Identify ways in which vast new sets of data on human beings can be collected, integrated, and analyzed to improve urban systems and quality of life while protecting confidentiality.

Provide both a theoretical and practical foundation which cities across the world can draw from in establishing their data access rules and data security procedures.

Authors

Sponsors

Robert Wood Johnson Foundation
NELL: Learning to Read the Web

Each day NELL:
1. Reads more facts from the web
2. Learns to read better than yesterday

Learning non-stop since January 2010

Fragment of NELL's ever-growing knowledge base, currently containing ~30 million beliefs
Microsoft Projects of Interest

- The Internet of (Orderly) Things
  - Exploitation of digital device information through cloud capture and storage of sensors in industrial or utility automation/monitor/monitor/wonder product information, etc.
  - Cloud ERP and historical correlations in ‘real-time’ to intervene in process optimization
  - Assimilation, Thematic, Predictive analytics for maintenance, electric and utility optimization
  - Industrial Process Automation: Create quality with control parameters and process monitoring data, optimization of the system, beyond device or process level

- The Internet of (Disorderly) Things
  - Unconventional, use case or vertical domain-specific, hard to search and analyze data.
  - Satellite images, nanoprograms, and sensor data (e.g., tweets, and opinion data; resume vetting, email analysis; website log files; genetics)
  - Healthcare: Genomic & image analysis, brain mapping, Pharma 3D clinical trial optimization
  - Labor Pool. Supply/demand matching

- Large Scale Analytical Computation
  - Interactive, exploratory work, leading to ‘automated production systems’ which incorporate models and algorithms
  - Reference data (medical sciences, economic data, government statistics, etc.) needed as part of the analytical process
  - Customization of portfolio performance attribution
  - On-demand, evaluate energy impact of design alternatives for architectural designs, structural analysis, kinematic simulation

SGI Solutions for Big Data

- Comprehensive
  - HPC
  - Cloud
  - Storage
  - Services

- Leadership
  - Speed
  - Scale
  - Efficiency

- Proven
  - NASA Earth Exchange powered by SGI ICEX
  - SGI UV2 -- 64 TB of Shared Memory, 1 Q/S
  - 8% of Hadoop Clusters run on SGI
  - 600 PBs shipped in 2012

www.sgi.com
Data Bridge: Solving the First & Last Mile Problems in Big Data

First Mile: Bringing the Long-tail of Science Data into Mainstream
Last Mile: Automation of Linking, Clustering and Discovering Heterogeneous Data
Last Mile: Discover Interesting Relationships
Data Bridge: NSF-funded Big Data Project
  – Apply Socio- metric Network Analysis (SNA) to data
  – Link through Multi-dimensional vectors
    • Similar to, but for data:
    • Explore Relationships between Data, Users, Resources, Methods, Workflows, ...
  – Architecture: Extensible, Highly Distributed, Plug & Play Algorithms, MyVector
    • Very loosely coupled Message-Oriented Middleware (using AMQP)
    • Built upon proven technologies: Integrated Rule Oriented Data Systems (iRODS), Dataverse Network

IBM Brings Big Data Skills to the Classroom and Workforce
Analyzing (Big) Data Boxes: Multi-way CS of Tensors

Motivating applications
- NELL @ CMU / Tom Mitchell
- Semantic analysis of brain fMRI
- Context-aware recommender systems (healthcare)

New analytical & computational tools
- Tensor does not fit in RAM – compress?
- Theorem: Latent factors sparse \( \Rightarrow \) can reduce down to \( LMN = O(F^{3/2}) \) and still recover the big red-blue-green factors exactly!
- Scalable coupled tensor-matrix computations (Hadoop)
- Reduced-rank tensor filtering

Myria: Foundations and Systems for Big Data Management

Q: Why are relational database so successful?
A: Because of Mathematical foundations + systems design

Myria = UW’s Big Data Management project
- Studies both mathematical foundations and systems’ design

- Mathematical foundations:
  - Communication + the new complexity parameter
  - Many models? How much data visualization per round?
  - Results: right rounding/approximation tradeoff for select-from-where queries
  - Exact steps: rounding/approximation tradeoff for skew, iteration, aggregates

- Systems design:
  - Old stuff = relational model, database, shared-nothing architecture
  - New features (models) = multi-way join operators, iteration, UDPs, LLDAs
  - New features (system) = cloud service, SLA, multi-tenants, fault tolerance


https://semanticommunity.info/A_NITRD_Dashboard/Making_the_Most_of_Big_Data
Updated: Thu, 27 Jun 2019 05:55:32 GMT
Powered by mindtouch
Projects of the NRC Board on Research Data & Info, and US CODATA

Analytical Approaches to Massive Data Analysis

Eli Upfal – Brown University

• Big data isn’t always the right data
• Data Mining/Machine Learning output isn’t always the significant, relevant output  How do we evaluate analytics results?
• Statistics vs. Machine Learning – formal statistical inference (p-value, confidence level) vs. computationally practical solutions
• Goal: Efficient data analysis with rigorous statistical guarantees

Recent work: Efficient and statistically sound tools for questions like:
• Is it a discovery or noise? Multi-hypothesis testing, FDR, VC-dim...
• Is the hypothesis wrong or the sample too small? Minimum sample size for detecting complex structures

Genomic applications

https://semanticommunity.info/A_NITRD_Dashboard/Making_the_Most_of_Big_Data
Updated: Thu, 27 Jun 2019 05:55:32 GMT
Powered by mindtouch™
Integrating Humans, Machines and Networks: A Global Review of Data-to-Decision Technologies

- Cognitive science
- Data analytics
- Decision science
- Machine learning
- Natural language processing
- Neuroscience
- Sensing and perception
- Software agents and multi-agent systems

A multinational review:
- Singapore
- Germany
- Turkey
- Etc.

Big Data Plus Cyber Security

Risk

Detection

Big Data

Agile

Psychosocial

WHITE HOUSE BIG DATA PARTNER WORKSHOP
National Weather Service for Infectious Disease

**National Weather Service for Infectious Disease**

**Predicting the Next Pandemic**

- Objective 1: Create Consortium to enhance multi-sector collaboration and examine how big data can be used to predict pandemics.
- Objective 2: Identify current capabilities, gaps, data requirements, and analytic needs necessary for development of a predictive capability; and
- Objective 3: Develop pilot project to demonstrate feasibility.

DOE Leadership Computing supports Big Data

**DOE Leadership Computing supports Big Data**

- Innovative and Novel Computational Impact on Theory and Experiment
  - Leadership Computing for Open Science (ALCF, OLCF)
  - Small number of projects (about 50) and users (about 800) with computationally intensive pair reviewed projects to advance science, speed innovation, and strengthen industrial competitiveness.
  - Also provides Big Compute for Big Data

Data Sharing and Metadata Curation: Obstacles and Strategies

**Data Sharing and Metadata Curation: Obstacles and Strategies**

Source: [Workshop Summary Report](https://semanticommunity.info/A_NITRD_Dashboard/Making_the_Most_of_Big_Data)

**Future strategies for managing scientific data and metadata for basic and applied research**

May 29, 2013
National Science Foundation, Room I-1235
4201 Wilson Blvd.
Introduction

This workshop was organized on behalf of the Big Data Senior Steering Group (BDSSG), an interagency body chartered by the White House Office of Science and Technology Policy (OSTP) and facilitated through the National Coordination Office for Information Technology and Networking R&D (NCO-NITRD). The goal was to have focused discussions on future strategies in data and metadata for basic and applied research; specifically, (a) how to better enable, encourage, and realize sharing of data, both across disciplinary divides and between "micro-silos" within research domains, and (b) how to acquire, manage, and curate metadata in order to ensure usability and comprehensibility of data over time and between disciplines.

Our intent was to bring together representatives from distinct data-intensive research domains who have been contributing to community-based solutions to data challenges. Key individuals from communities such as the Materials Genome Initiative, Space Weather, Global Climate, Environmental Health, DataOne, iRODS, and the Research Data Alliance participated.

The workshop was comprised of three main sessions. The first session featured five presentations from practitioners who focus on a particular domain; the second session included three groups that focus on trans disciplinary data; and the third session was an open discussion focused around several questions that had been distributed to the participants in advance.

Executive Summary

Attendees were asked to consider the following five questions.

1. What metadata, and what kinds of metadata management, are needed to enable re-use of data, both across domains and across silos within domains?
2. How can we incentivize researchers and providers to curate their data, organize it with useful metadata, and make it publicly available?
3. Maximum impact of data occurs when analytics make use of all available relevant data; how can analytics developers be challenged to make this standard practice?
4. What are the data ownership and personal identifiable information issues (obstacles/solutions) that can be addressed in this context?
5. What are the top two data/metadata problems you would like to solve?

There were at least four areas of agreement from our discussions:

• Active data stewardship/curation adds value and is needed at some level. However, cost is a major issue. There is no funding model to support the resources needed, and no way to assess the value of data management compared to, e.g. new research grants.
• Exclusively top-down solutions are not desired; but the correct balance between grass-roots vs. "middle-out" initiatives is unclear.
• There is a need for easy-to-use tools for metadata creation, improvement, and workflows that incorporate good data practices.
• Funding agencies can provide incentives for researchers to share data. For example, applicants could receive credit for making their data more readily accessible through the use of community best practices for sharing; funded researchers could be required to use their research field’s metadata standards.
Discussion and Presentation Summary

The following summary is intended for the workshop participants and the members of the BDSSG and is not meant to be a complete review of the subject. It is comprised of summary notes and links to the presenter’s slides and video.

Practitioners’ Perspectives

Moderated by Robert Chadduck, NSF

- **DataOne** (DataNet Observational Network for the Earth) - Rebecca Koskela, University of New Mexico [slides; video]
  - *Purpose*: to promote data discovery in earth/environmental sciences.
  - *Method*: Three major nodes (CA, NM, ORNL) fed by member nodes. Becoming a member node gives more visibility to your data.
  - *Provide*: investigator toolkit; process to align diverse metadata; index metadata for the search API; Tools such as DataUP, OneShare, and Dryad to help researchers improve data practices, create metadata, help with uploading, repositories and DOIs.
  - *Working on*: Semantic mediation, provenance, and automated annotation
  - *Observations*: Of the scientists they work with:
    - 80% want to share data, but only 6% share all of their data
    - have almost no metadata standards
    - most use Excel spreadsheets

- **DFC/iRODS** (DataNet Federation Consortium/innovative rules oriented data system)- Reagan Moore and Mary Whitton, RENCI [slides; video]
  - *Purpose*: Provide a federated collaboration environment that supports reproducible data-driven research.
  - *Provide*: Mechanisms to enable interoperability and allow domains and services to interact. Not just metadata but the procedures used to create the data product: procedures for data acquisition, data management, automation of the application of domain knowledge; policies for data control. iRODS policy-based data management.
  - *Working on*: Encapsulation of domain knowledge for accessing domain repositories, analyzing domain data sets, and managing domain data products. Application of virtualization mechanisms that manage metadata properties and the processes to derive the metadata.

- **NIST/ITL/MML** (NIST Information Technology Laboratory and Material Measurement Lab)- Mary Brady, Ram Sriram, NIST ITL, and Jim Warren, Carelyn Campbell, NIST MML [slides; video] My Note: These are one in the same
  - *Purpose*: To facilitate the Material Genome Project by enabling data exchange, ensuring data quality, and establishing new methods and metrologies.
  - *Provide*: Developed repositories and other necessary infrastructure. Currently moderated submission but working toward more automation. Standing up office of data and informatics at NIST, developing universal identifiers and ontologies for materials development.

- **NCN/nanoHUB** (Network for Computational Nanotechnology)- Gerhard Klimeck, Purdue [slides; video]
  - *Purpose*: Resource for the use of the Nanotechnology Community of Researchers
  - *Provide*: Simulation tools, collaboration tools; resources to teach and learn such as nanoHUB-U, courses, seminars, and tools to share and publish tools and research.
  - *Observations*:
    - Perceived myths:
      - You can't use research codes for education, you must write your own,
• Building user interfaces is too hard, you must rewrite for the web,
• There are no incentives to share and no end-to-end science cloud possible.
• Observations regarding these myths: Large development collaborations that serve large number of users = predictable success.
• Criteria for the success of a science gateway:
  • Outstanding science,
  • Commitment to dissemination,
  • Technology for dissemination,
  • Tech transfer process (i.e. people),
  • Understand the stakeholders,
  • Open assessment, and
  • A business model. Consider the iPad…may not be as capable as a typical desktop, but it’s much more useable.

• **BMIR** (Stanford Center for Biomedical Informatics Research) – *Mark Musen, Stanford slides; video*
  ◦ **Purpose**: Research to improve the exchange of health information
  ◦ **Current State**:
    • Research data:
      • [BioSharing Initiative](#) – tried to provide a path through all the data and metadata policies, standards, databases.
      • [BioDBcore](#) – uniform description of public biological databases
      • Minimal Information About a Microarray Experiment Initiative – grass roots standard that is now adopted by some organizations. Is leading to a markup language and ontology. Many different “minimal information” checklists under the [MIBBI](#) portal, all grass-roots efforts
    • Clinical data:
      • [HL7](#) Organization’s Reference Information Model has had limited adoption (too complex).
  ◦ **Observations**:
    • Development of meaningful use criteria is a necessary first step (also the conclusion of a PCAST report).
    • There must be progress made toward a robust exchange of health information.
    • We need a universal exchange language and an IT infrastructure to support it, but this is not what the vendors involved in HL7 are envisioning.
    • Many metadata solutions have been met with outright hostility.

• **Open Discussion – Practitioners’ Perspectives**
  ◦ **“What keeps you up at night?”**
    • Lack of strong incentives,
    • Attribution to individuals is an important incentive,
    • There are concerns about sharing standards, ensuring quality, requiring open source, that need to be addressed,
    • Data sharing is not in the workflow and needs to be,
    • Metadata collection and generation is not scalable to meet the needs,
    • Tools/solutions must be pragmatic and integrated into the workflow,
Can we do an “overarching ontology”? Perhaps the best we can do is start with small ontologies as a foothold into crossing domains,

- Administrative metadata is standard, but descriptive and provenance is not standard across domains,
- Uses for metadata include provenance and curation, description, and state information
- The context in which the data was acquired is crucial for getting out of the silo.

Trans-disciplinary Community Perspectives:

Moderated by Alan Hall, NOAA, and Jon Petters, AAAS Fellow at DOE

- **RDA** (Research Data Alliance) - *Fran Berman, Rensselaer Polytechnic* slides; video
  - *Purpose:* to build social, organizational, technical infrastructure to reduce barriers to data sharing and accelerate development of coordinated global infrastructure.
  - *Method:*
    - Working groups work for 12-18 months to build targeted pieces of infrastructure. Interest groups include agricultural data, big data analytics, legal, etc. WG examples are: Persistent Identifier Information Types, Data Type Registries, Data Foundation and Terminology, Practical Policy (latter 2 pending).
    - 45 countries participating, over 700 participants, US =31% of participants. Academics = 61%, private sector 21%. US leadership considered strategically important.

- **EarthCube** *Joel Cutcher-Gershenfeld, University of Illinois* slides; video
  - *Purpose:* To study institutional context and the impact of “culture” in the context of a community such as EarthCube.
  - *Method:*
    - Reaching out and hearing the voice of the customer. This was a key piece in the development of EarthCube. Results of stakeholder survey = Everybody thinks using disparate data and tools is important but very hard to do.
    - Power is not in pushing on the drivers but in lowering the barriers, ex: if you lower a barrier (like lack of credit) then you create an incentive (credit).
    - Engage cross-cultural conversations; Culture eats strategy for breakfast.
    - Be wary of building a great system that no one uses.

- **Interagency Perspectives** - *Ted Haberman, HDF Group* slides; video
  - *Observations:*
    - We are in an "era of ferment" where there are risks, uncertainty, waste, lack of interoperability. Will reach a point where there is a selection of options, focusing of energy, and a convergence of communities.
    - A lot of the "tools" we talk about are basically portals. Instead, tools for creators of metadata are needed--rubrics for evaluation, views, connections, development, evolution, and documentation.
    - There is a need for a community support environment (e.g. a Documentation Consortium).

- **Open Discussion- Community**
  - *“What are the successes and worries?”*
    - Success: There is energy around RDA. Worries: Stewardship and economics of stewardship. How to store it and pay for it is a multi-sector issue.
    - Success: Groups are becoming more intentional and focused. Worries: “It takes internal alignment to get lateral alignment.” There are risks that you can build it but nobody will come.
• Success: Lots of resources going into things. Worries: Resources are not being used to contribute to the big picture (ferment) e.g. lots of documents (words being produced), but converting them into actions is a worry.

• It is hard to do something new and still maintain credibility in your "home domain".

  ◦ "How do we focus on building cultural infrastructure and removing barriers?"
    • Similar to academic cultures in universities, you have to incentivize the desired behaviors in a mainstream way. Data sharing is dis-incentivized by competition.
    • Behavior changes culture but only very slowly.
    • Find the "positive deviants", people who are already doing things "right" and share their successes.

  ◦ "We need to keep data that is valuable, and discard that which is irrelevant. Who is responsible for paying for and doing this kind of curation?"
    • Reference to Sustainable Economics for a Digital Planet and other documents on Fran Berman’s website. We have to understand that every dataset has multiple stakeholders. Who takes it? Who keeps it? Who manages it?
      • It is time to develop reasonable policies; e.g. what's the value of the data and how hard would it be to regenerate it? The same infrastructure is not necessary for all types of data.
      • Some datasets are interesting only in the context of a paper.
      • In some cases, the cost of regeneration could be huge compared to the cost of curation.
    • There are organizations in the EU that don't exist in the US. There needs to be outreach to the science communities to let them know that a larger community is forming.
    • Identify the leaders, fund them, and make exemplars of them.

Open Discussion: Barriers and Opportunities:

Moderated by Peter Lyster, NIH, and Mark Suskin

• “What does "metadata" mean?”
  ◦ Metadata:
    • Is what you need to ensure that the person you will never meet will not reach incorrect conclusions by using your data.
    • Is contextual and implies active curation.
    • Is used but not talked about: In smart laboratories this information is standardized for at least certain kinds of experiments. But the idea that the metadata you save is supposed to plug into a global infrastructure isn't talked about.
    • Is expensive: nanoHUB is spending 60% of its budget on content stewardship i.e. curation. It requires a PhD level person who can interact with colleagues.
    • Has few but expensive experts: who is it that wants to learn best practices in data curation? E.g. Kirk Borne's efforts at GMU, teaching data practices to their astronomy students.
    • Has few built-in incentives: Curation doesn't lead to scientific publication.

• What does it mean to succeed in any of these areas? Can we lay out clear-cut desired outcomes? What should the Government do and not do?
  ◦ Start with what you have already invested in and build on it.
  ◦ High level abstractions that make searches simple is a lesson from business data management. For example “faceted” classification has been adopted and used for the Earth System Grid Portal. It could be developed and used for research funded by multiple agencies (see Habermann’s presentation).
○ Develop a grand challenge to develop a cost model. Are we willing to sacrifice research grants for better curation of datasets?
○ Create an environment where there is room for an entrepreneur to do something new.

• In response to the question: What are the top 2 metadata problems?
  ○ The Lack of:
    ▪ Success stories to demonstrate return on investment.
    ▪ Modeling from Government agencies who:
      ▪ Do not collaborate or use international metadata standards effectively.
      ▪ Do not provide clarity for what will be supported by the public sector and how, and what should be supported elsewhere.
  ○ Agreement on:
    ▪ Scientific ontologies that allow categorization at the right level of abstraction to facilitate the creation of metadata tools.
    ▪ Templates and standards so everybody knows what they are supposed to deliver.
    ▪ Persistent identifiers for everything (to build trust).
    ▪ Semantic standards.
    ▪ Consistent IP rights across data in the US e.g. credit, provenance, citation.
  ○ Innovative Tools:
    ▪ That IMPROVE metadata e.g. reduce uncertainties that cause errors
    ▪ For data documentation, that points to available standards.
    ▪ That fit data curation into the workflow
  ○ A Community to:
    ▪ Share best practices and soft knowledge e.g. workflow libraries
    ▪ Help establish the identity of existing collections.
    ▪ Find stable homes for valuable data.
    ▪ Teach data literacy e.g. educating researchers in preserving and sharing data
    ▪ Provide a platform for scholarly communication that isn't publishing a paper but communicating inside a network, "adding data points rather than producing inaccessible works."
  ○ Incentives that:
    ▪ Treat data as a 1st class publication if it's fully integrated with context.
    ▪ Change the value perception of metadata for the PI.
    ▪ Grade scientists on how well they cite their data.
    ▪ Find no-cost policy changes that enable credit to accrue to data creators.

Summary and Wrap-up:

Tom Statler, NSF slides

With Big Data comes big risk: risk of reaching incorrect conclusions (through misunderstanding, misuse, or abuse of data), risk of data investment losing value, risk of data becoming unusable. Metadata is the essential information needed to minimize these risks. Metadata curation is managing these risks and accepting them where appropriate. Standards and practices developed for domain-specific needs are just starting to interact. The hazard of a top-down unification of
standards across domains is that it can appear to lower barriers while being doomed to internal fracturing; the sociological problem may be as hard as the technological one.

Johns Hopkin University's Grant Reviewer's Guide was offered as just one example of policy guidance that can be cost effective even in the short term: http://dmp.data.jhu.edu/resources/gr...viewers-guide/.

The workshop concluded with a challenge to the participants, “What are YOU going to do?”

- “Talk with my program officer about highlighting dataset developments.”
- “DOE Office of Science is giving guidance to PIs and POS about using data management as part of evaluation. For attribution and citation, it’s harder but on the list.”
- “Helping to organize meetings and sessions. All of this stuff for data also applies to software.”
- “NCO will vigorously support the BDSSG and domain group.”
- “Get outside of one's own portfolio.”
- “My WG will deliver a prototype of a digital object registry.”
- “Bring discussions back to NIH, to three workshops.”

Workshop Summary and Wrap-Up

Workshop Summary and Wrap-Up

Tom Statler, NSF
With Big Data comes big risk

- Risk of reaching incorrect conclusions
- through misunderstanding of data
- through misuse or abuse of data
- Risk of data investment losing value
- Risk of data becoming unusable

- Metadata is the essential information needed to minimize these risks.
- Metadata curation is managing these risks and accepting them where appropriate.

Aligning metadata across pre-aligned domains?

Standards & practices developed for domain-specific needs are just starting to interact.
The hazard of a top-down unification of standards across domains is that it can appear to lower barriers while being doomed to internal fracturing.
The sociological problem is as hard as than the technological problem

- Understanding needs of stakeholders
- Sharing can disincentivize competition
- Difficult for researchers to contribute in a new field and yet retain credibility in their home disciplines.
- “Culture eats strategy for breakfast.”
- “It takes internal alignment to get horizontal alignment.”
- Can we measure the value of a dataset?

Areas of Agreement?

- Active data stewardship/curation adds value and is needed at some level; but we have no funding model to support these people and no way to measure the value compared to, e.g., new research grants.
- We’re not looking for exclusively top-down “we’re from the government and we’re here to help” solutions; but we don’t know what the correct balance of grass-roots vs. “middle-out” initiatives is.
- There is a need for easy-to-use tools for metadata creation, improvement, and workflows that incorporate good data practices.
So what next?

What should federal agencies do (at “zero cost”)?

What should communities do internally?

What should universities/academic researchers do?

What should the private sector do?

Keyword: DO (On-the-job training for everybody!)

Participant List

Bell, Randy, DOE/NNSA
Berman, Fran, RPI/RDA
Biven, Laura, DOE/SC
Brady, Mary, NIST
Campbell, Carelyn, NIST/MML
Cutcher-Gershenfeld, Joel, UI
de Groot-Lief, Christina, NOAA
Eigen, Ana, DOT
Gupta, Amarnath, UCSD
Habermann, Ted, The HDF Group
Hagan, Don, CENDI/NTIS
Hamilton, Carol, RTI
Klimeck, Gerhard, Purdue
Kolker, Eugene, SCRI
Koskela, Rebecca, DataONE
Kuznetsova, Maria, NASA
Larkin, Jennie, NIH
Lewis, Suzanna, BBOP
Maddox, Marlo, NASA
Martone, Maryann, UCSD
McDermott, Michael, USGS
Moore, Reagan, RENCI
Musen, Mark, Stanford
Rindflesch, Tom, NIH/NLM
Scott, John Henry, NIST
Shoshani, Arie, LBL
Smith, Barry, NCOR
Szalay, Alex, JHU
Tilmes, Curt, NASA/USGCRP
Vieglais, Dave, UK/DataONE
Viereck, Rodney, NOAA
Ward, Charles, AFRL
Warren, Jim, NIST/MML
Whitton, Mary, RENCI
Wilbanks, John, NCO/NITRD

Organizing Committee
Bristol, Sky, USGS
Chadduck, Bob, NSF
Dearry, Allen, NIH
Grumbling, Emily, NSF
Hall, Alan, NOAA
Lee, Tsengdar, NASA
Lyster, Peter, NIH
Glossary of Acronyms

AFRL: Air Force Research Laboratory
BBOP: Berkeley Bioinformatics Open-source Projects
CENDI/NTIS: An interagency working group of senior scientific and technical information (STI) managers/ National Technical Information Service
DataONE: Data Observation Network for Earth
DOE/NNSA: Department of Energy/National Nuclear Security Administration
DOE/SC: Department of Energy/Department of Science
DOT: Department of Transportation
HDF: Hierarchical Data Format
JHU: Johns Hopkins University
LBL: Lawrence Berkeley National Laboratory
NASA: National Aeronautic and Space Administration
NASA/USGCRP: National Aeronautic and Space Administration/United States Global Change Research Program
NCOR: National Center for Ontological Research
NCO/NITRD: National Coordination Office/Networking and Information Technology Research and Development
NIH: National Institutes of Health
NIH/NIGMS: National Institutes of Health/National Institute of General Medical Science
NIH/NLM: National Institutes of Health/National Library of Medicine
NIST: National Institute of Standards and Technology
NIST/MML: National Institute of Standards and Technology/Material Measurement Laboratory
NOAA: National Oceanic and Atmospheric Administration