Data Science Symposium 2013

Story

NIST seeks to bring rigor to data science

Agenda for the NIST Data Science Symposium 2014

Day 1 - March 4, 2014
   Arrival
   Welcome & Opening Remarks
   The NIST Data Science Program and Symposium Objectives
   Challenges and Gaps in Data Science Technologies
   Data and Use Cases for Data Science Research
   Lunch
   Data Science Poster Session
   Data Science Benchmarking and Measurement

Day 2 - March 5, 2014
   Arrival
   Keynote Address
   Breakout Sessions
   Lunch
   Government Data Science R&D Panel
   Breakout Reports

Abstract

Story

Slides

Slide 1 What a Data Scientist Does and How They Do It
Slide 2 Overview
Slide 3 Brief Bio
Slide 4 The State of Federal Data Science: My Activities
Slide 5 The State of Federal Data Science: What is Big Data?
Slide 6 The State of Federal Data Science: My Big Data Matrix
Slide 7 Data Science Team Example: Open Government Data Manager
Slide 8 Data Science Team Example: Chief Data Science Officer
Slide 9 Data Science, Data Products, and Data Stories Venn Diagrams
Slide 10 Data Science DC Community Posts: Cloud SOA Semantics and Data Science Conference
Slide 11 Data Science DC Community Posts: Event Review: Confidential Data
Slide 12 Data Science DC Community Posts: Data Science DC In Review
Slide 13 Data Science DC Community Posts: Data Science DC Data in Spotfire Web Player
Slide 14 Data Science DC Community Posts: Data Science, Data Products, and Predictive Analytics 1
Slide 15 Data Science DC Community Posts: Data Science, Data Products, and Predictive Analytics 2
Slide 16 Data Science DC Community Posts: Data Science, Data Products, and Predictive Analytics 3
Slide 17 NAS Report on Frontiers in Massive Data Analysis
Slide 18 NAS Report on Frontiers in Massive Data Analysis: Application to the NIST Big Data Public Working Group
Slide 19 NAS Report on Frontiers in Massive Data Analysis: Definition
Slide 20 NAS Report on Frontiers in Massive Data Analysis: Taxonomy
Slide 21 NAS Report on Frontiers in Massive Data Analysis: Reference Architecture
Slide 23 NAS Report on Frontiers in Massive Data Analysis: Roadmap
Slide 24 NAS Report on Frontiers in Massive Data Analysis: Make a Technology Roadmap Project a Data Project
Slide 25 Graph Databases and the Semantic Web: New Book and Tutorial on Neo4j
Slide 26 Graph Databases and the Semantic Web: My Talking Points
Slide 27 Graph Databases and the Semantic Web: Knowledge Base of NIST Symposium and BD-PWG Use Cases
Slide 28 Graph Databases and the Semantic Web: Semantic Web Linked Data for Triples
Slide 29 Graph Databases and the Semantic Web: Spotfire Cover Page
Slide 30 Graph Databases and the Semantic Web: NIST Big Data Public Work Group Use Cases and Knowledge Bases in Spotfire
Slide 31 Semantic Medline – YarcData Graph Appliance Application for Federal Big Data Senior Steering WG: Graphs and Traditional Technologies
Slide 32 Semantic Medline – YarcData Graph Appliance Application for Federal Big Data Senior Steering WG: The YarcData Approach
Slide 33 Semantic Medline – YarcData Graph Appliance Application for Federal Big Data Senior Steering WG: Semantic Medline at NIH-NLM
Slide 34 Semantic Medline – YarcData Graph Appliance Application for Federal Big Data Senior Steering WG: Bioinformatics Publication
Slide 35 Semantic Medline – YarcData Graph Appliance Application for Federal Big Data Senior Steering WG: Semantic Medline Database Application
Slide 36 Semantic Medline – YarcData Graph Appliance Application for Federal Big Data Senior Steering WG: Work Flow
Slide 37 Semantic Medline – YarcData Graph Appliance Application for Federal Big Data Senior Steering WG: Predication Structure for Text Extraction of Triples
Agenda

Full High-Level Use Case Descriptions

1. Introduction

2. Use Case Summaries

Government Operation

2.1 Census 2010 and 2000 – Title 13 Big Data; Vivek Navale & Quyen Nguyen, NARA

2.2 National Archives and Records Administration Accession NARA, Search, Retrieve, Preservation; Vivek Navale & Quyen Nguyen, NARA

2.3 Statistical Survey Response Improvement (Adaptive Design); Cavan Capps, U.S. Census Bureau

2.4 Non-Traditional Data in Statistical Survey Response Improvement (Adaptive Design); Cavan Capps, U.S. Census Bureau

Commercial

2.5 Cloud Eco-System, for Financial Industries (Banking, Securities & Investments, Insurance) transacting business within the United States; Pw Carey, Compliance Partners, LLC

2.6 Mendeley – An International Network of Research; William Gunn, Mendeley

2.7 Netflix Movie Service; Geoffrey Fox, Indiana University

2.8 Web Search; Geoffrey Fox, Indiana University

2.9 IaaS (Infrastructure as a Service) Big Data Business Continuity & Disaster Recovery (BC/DR) Within a Cloud Eco-System; Pw Carey, Compliance Partners, LLC

2.10 Cargo Shipping; William Miller, MaCT USA

2.11 Materials Data for Manufacturing; John Rumble, R&R Data Services

2.12 Simulation driven Materials Genomics; David Skinner, LBNL

Defense

2.13 Large Scale Geospatial Analysis and Visualization; David Boyd, Data Tactics

2.14 Object identification and tracking from Wide Area Large Format Imagery (WALF) Imagery or Full Motion Video (FMV) – Persistent Surveillance; David Boyd, Data Tactics
2.15 Intelligence Data Processing and Analysis; David Boyd, Data Tactics

Healthcare and Life Sciences

2.16 Electronic Medical Record (EMR) Data; Shaun Grannis, Indiana University

2.17 Pathology Imaging/digital pathology; Fusheng Wang, Emory University

Figure 1: Examples of 2-D and 3-D pathology images

Figure 2: Architecture of Hadoop-GIS, a spatial data warehousing system over MapReduce to support spatial analytics for analytical pathology imaging

2.18 Computational Bioimaging; David Skinner, Joaquin Correa, Daniela Ushizima, Joerg Meyer, LBNL

2.19 Genomic Measurements; Justin Zook, NIST

2.20 Comparative analysis for metagenomes and genomes; Ernest Szeto, LBNL (Joint Genome Institute)

2.21 Individualized Diabetes Management; Ying Ding, Indiana University

2.22 Statistical Relational Artificial Intelligence for Health Care; Sriraam Natarajan, Indiana University

2.23 World Population Scale Epidemiological Study; Madhav Marathe, Stephen Eubank or Chris Barrett, Virginia Tech

2.24 Social Contagion Modeling for Planning, Public Health and Disaster Management; Madhav Marathe or Chris Kuhlman, Virginia Tech

2.25 Biodiversity and LifeWatch; Wouter Los, Yuri Demchenko, University of Amsterdam

Deep Learning and Social Media

2.26 Large-scale Deep Learning; Adam Coates, Stanford University

2.27 Organizing large-scale, unstructured collections of consumer photos; David Crandall, Indiana University

2.28 Truthy: Information diffusion research from Twitter Data; Filippo Menczer, Alessandro Flammini, Emilio Ferrara, Indiana University

2.29 Crowd Sourcing in the Humanities as Source for Big and Dynamic Data; Sebastian Drude, Max-Planck-Institute for Psycholinguistics, Nijmegen The Netherlands

2.30 CINET: Cyberinfrastructure for Network (Graph) Science and Analytics; Madhav Marathe or Keith Bisset, Virginia Tech

2.31 NIST Information Access Division analytic technology performance measurement, evaluations, and standards; John Garofolo, NIST

The Ecosystem for Research

2.32 DataNet Federation Consortium DFC; Reagan Moore, University of North Carolina at Chapel Hill

Figure Policy-based Data Management Concept Graph (iRODS)

2.33 The ‘Discinnet process’, metadata <-> big data global experiment; P. Journeau, Discinnet Labs

2.34 Enabling Face-Book like Semantic Graph-search on Scientific Chemical and Text-based Data; Talapady Bhat, NIST

2.35 Light source beamlines; Eli Dart, LBNL

Astronomy and Physics

http://semanticommunity.info/Data_Science/Data_Science_Symposium_2013

Updated: Wed, 23 Sep 2015 02:59:49 GMT
Powered by mindtouch
2.36 Catalina Real-Time Transient Survey (CRTS): a digital, panoramic, synoptic sky survey; S. G. Djorgovski, Caltech

Figure Catalina Real-Time Transient Survey (CRTS): a digital, panoramic, synoptic sky survey

2.37 DOE Extreme Data from Cosmological Sky Survey and Simulations; Salman Habib, Argonne National Laboratory; Andrew Connolly, University of Washington

2.38 Large Survey Data for Cosmology; Peter Nugent LBNL

2.39 Particle Physics: Analysis of LHC Large Hadron Collider Data: Discovery of Higgs particle; Michael Ernst BNL, Lothar Bauerdick FNAL, Geoffrey Fox, Indiana University; Eli Dart, LBNL

Figure 1 The LHC Collider location at CERN
Figure 2 The Multi-tier LHC computing infrastructure

2.40 Belle II High Energy Physics Experiment; David Asner & Malachi Schram, PNNL

Earth, Environmental and Polar Science

2.41 EISCAT 3D incoherent scatter radar system; Yin Chen, Cardiff University; Ingemar Häggström, Ingrid Mann, Craig Heinselman, EISCAT Science Association

2.42 ENVRI, Common Operations of Environmental Research Infrastructure; Yin Chen, Cardiff University

Figure 1: ENVRI Common Subsystems
Figure 2: Architectures of the ESFRI Environmental Research Infrastructures (A) ICOS Architecture
Figure 2: Architectures of the ESFRI Environmental Research Infrastructures (B) LifeWatch Architecture
Figure 2: Architectures of the ESFRI Environmental Research Infrastructures (C) EMSO Architecture
Figure 2: Architectures of the ESFRI Environmental Research Infrastructures (D) Eura-Argo Architecture
Figure 2: Architectures of the ESFRI Environmental Research Infrastructures (E) EISCAT 3D Architecture

2.43 Radar Data Analysis for CReSIS Remote Sensing of Ice Sheets; Geoffrey Fox, Indiana University

Figure 1: Typical Radar Data after analysis
Figure 2: Typical flight paths of data gathering in survey region
Figure 3. Typical echogram with Detected Boundaries. The upper (green) boundary is between air and ice layer while the lower (red) boundary is between ice and terrain

2.44 UAVSAR Data Processing, Data Product Delivery, and Data Services; Andrea Donnellan and Jay Parker, NASA JPL

2.45 NASA LARC/GSFC iRODS Federation Testbed; Brandi Quam, NASA Langley Research Center

2.46 MERRA Analytic Services MERRA/AS; John L. Schnase & Daniel Q. Duffy, NASA Goddard Space Flight Center

Figure Typical MERRA/AS Output

2.47 Atmospheric Turbulence - Event Discovery and Predictive Analytics; Michael Seablom, NASA HQ
The NIST Data Science Symposium 2014

This symposium was scheduled for last November 18-19, 2013, but the Federal government was just coming to work after a historic shutdown over budget matters.

The agenda says we will hear about:

- The NIST Data Science Program
- Data Science Technologies
- Data Science Data and Use Cases
- Data Science Posters
- Data Science Benchmarking & Performance Measurement
- Government Data Science R&D
- Big Data Analytics for Smart Manufacturing Systems

The last topic is a new one to me and would seem to fit more in a Big Data Symposium.

The real question seems to be how much real data science will we actually hear in the presentations and see in the posters?

My quick answer is very little because this was more about NIST learning what data science is and its possible intersection with their specialty, measurement science, so it seems to ultimately about metrics for data science because standardizing metrics is NIST's mission. The prime example is when you see the wall clock in their conference room you know that is the standard for time!

At the first break after the opening session, I was interviewed by Joab Jackson with the IDG News Service on my thoughts and I just found them from a Google search appearing in at least four publications! See below for the full story. My quote was:
"A lot of companies are finding that they thought they were getting data science when they purchased Hadoop, but then they have to hire a data scientist to really do something useful with it," said symposium attendee Brand Niemann, director and senior enterprise architect at the data management consulting firm Semantic Community.

I walked around to look at the 38 posters whose titles and descriptions are in the Proceedings. Actually I assume there will be more proceedings based on the Breakout Session Reports and Next Steps at the end of the Symposium.

I found the Breakout Sessions and Focus Groups to be most informative as follows:

- Focus Group on HCI and Data Science (Sharon Laskowski)
  - The handout contained six questions (e.g. What are the current major data science needs for your users/customers? What role has HCI (human-computer interaction) played in addressing these needs? My response was Data Stories/Data Products that I do as a Data Scientist/Data Journalist and referred to my poster on the wall for more details.

- Datasets and Use Cases for Data Science Research (Ashit Talukder)
  - I have submitted a Use Case to the NIST Big Data Working Group (that has a standard template for their 50 or so use cases) and suggested to NIST’s Dr. Chris Greer and Wo Chang that they crowd-source Data Science on NIST’s key data sets that can be made public to benefit both the NIST Big Data and Data Science activities.

- Automated Metadata and Ontologies for Heterogenous Data (Anne Plant)
  - We are in phase three of this (Phase one was Metadata Clearinghouses because the data was not online, Phase two was actual data was online but the metadata was insufficient, and Phase 3 Data Science where the data scientist creates his own metadata and data products described in a Data Story that the public and technical people can understand and reuse. See three cases studies for Todd Park (Health Datapalooza, Heritage Health Prize, and IOM-CMS Data Science) and Data Science Makes Data More Important Than Code and Ontology.

My comments on some of the key speakers are as follows:

- John Garofolo - Senior Advisor for Program Development, IAD (NIST); Charles Romine - Director of the Information Technology Laboratory (NIST); Ashit Talukder - Chief, Information Access Division (NIST); and Craig Greenberg - Project Leader, Multimodal Information Group (NIST)
  - Are all learning about Data Science and seem to be unaware of the NIST Big Data Working Group efforts that includes Data Science

- Doug Cutting - Chief Architect (Cloudera)
  - This was a commercial for Cloudera and their new Enterprise Data Hub, not Data Science and is the basis for my quote above. After you use their framework and open source software, what do you actually do with the data?

- Haesun Park - Professor & Executive Director, Center for Data Analytics (Georgia Tech)
  - This talk should have used the NAS Report on Frontiers in Massive Data Analysis as I pointed out in my comments to him. He referenced Wikipedia Visual Analytics

- Eric Frost - Director, Viz Center (San Diego State) did not show and was replaced by Volker Markl - Professor and Chair of The Database Systems & Information Management Group (TU Berlin) - Data Science Research from the Database Management Systems Perspective
  - See http://stratosphere.eu and GitHub

- Jeanne Holm - Evangelist (Data.gov) did not show up and was replaced by Phil Ashock. Data.gov Chief Architect
Data.gov should be named DataCatalog.gov and Phil did not speak about the title: Datasets that Enable Rigorous Data Science Research. The Federal: Big Data Working Group Meetup does address both Datasets and Data Science that Enables Rigorous Data Science Research and Data Science for Business.

- Stanley Ahalt - Director, Renaissance Computing Institute (University of North Carolina)
  - I do not remember anything remarkable about this presentation

- Michael Hurley - Technical Staff (MIT Lincoln Laboratory)
  - Michael gave a talk that caused me to raise the issue about sharding big data across multiple computers and concluding that the average of the individual results are the same as the result from being able to run the complete big data in memory

- Chelsey Richards - Deputy Director for Public Health Scientific Services (CDC)
  - Chelsey, who directs the National Center for Health Statistics (NCHS), gave the best presentation in my opinion and considers TBs as big data

- Thomas Karl - Director, National Climatic Data Center (NOAA), etc.
  - I did not attend the Data Science Benchmarking and Measurement Session because I was still involved in the Poster Session

- Dr. Francine Berman - Chair, Research Data Alliance / U.S.
  - I suggested she become aware of the large Data Community and Data Science Community locally and internationally.

- Christopher White - Program Manager, XDATA Program (DARPA)
  - See DARPA Open Catalog

Ashit Talukder - Chief, Information Access Division (NIST) summarized the first day as follows:

- About 350 of the 500 registered showed up which was good considering the adverse weather
- There was much diversity and cross-cutting content
- A consensus seemed to be forming
- Academia is moving forward
- NIST needs to measure the data science processes
- NIST has an initial taxonomy to review and add to
- Data Science fits with NIST’s other programs for cloud, mobility, and big data

MORE TO FOLLOW

NIST seeks to bring rigor to data science

Sources: [http://news.idg.no/cw/art.cfm?id=96B4895B-D8EE-C894-2A08AF664F89C5CD](http://news.idg.no/cw/art.cfm?id=96B4895B-D8EE-C894-2A08AF664F89C5CD)


NIST plans to develop a framework to help different industries better understand and use big data

Joab Jackson (IDG News Service) 04 March, 2014 22:07

Pak Chung Wong, chief scientist at the Pacific Northwest national Laboratory, explains the challenges of working with enormous data sets, as the NIST data science symposium

The U.S. National Institute of Standards and Technology (NIST) wants to bring some metrics and rigor to the nascent but rapidly growing field of data science.

The government agency is embarking on a project to develop by 2016 a framework that can be used by all industries to understand how to use, and measure the results from data science, and big data projects.

NIST, an agency of the U.S. Department of Commerce, is holding a symposium on Tuesday and Wednesday at its Gaithersburg, Maryland, headquarters with big data specialists and data scientists to better understand the challenges around the emerging discipline.

"Data science is pretty convoluted because it involves multiple data types, structured and unstructured," said event speaker Ashit Talukder, NIST chief for its information access division. "So metrics to measure the performance of data science solutions is going to be pretty complex."

Starting with this symposium, the organization plans to seek feedback from industry about the challenges and successes of data science and big data projects. It then hopes to build a common taxonomy with the community that can be used across different domains of expertise, allowing best practices to be shared among multiple industries, Talukder said.
While computer-based data analysis is nothing new, many of the speakers at the event talked about a fundamental industry shift now going on underway with data analysis.

Doug Cutting, who originally created the Hadoop data processing platform noted that what made Hadoop unique is that it took a different approach to working with data. Instead of moving the data to a place where it can be analyzed -- an approach used with data warehouses -- the analysis takes place where the data is stored itself.

"You can't move [large] data sets without major performance penalties," Cutting said. Since its creation in 2005, Apache Hadoop has set the stage for storing and analyzing data sets so large that they can not fit into a standard relational database, hence the term "big data."

As these data sets grow larger, the tools for working with them are changing as well, noted Volker Markl, a professor and chair of the database systems and information management group at the Technische Universität Berlin.

"Data analysis is becoming more complex," Markl said. As a discipline, data science is challenging in that it requires both understanding the technologies to handle the data, such as Hadoop and R, as well as the statistics and other forms of mathematics needed to harvest useful information from the data, Markl said.

"A lot of companies are finding that they thought they were getting data science when they purchased Hadoop, but then they have to hire a data scientist to really do something useful with it," said symposium attendee Brand Niemann, director and senior enterprise architect at the data management consulting firm Semantic Community.

Another emerging problem with data science is that it is very difficult to maintain a data analysis system over time, given its complexity. As the people who developed the algorithms to analyze data move on to other jobs or retire, an organization may have difficulty finding other people to understand how the code works, Markl said.

Another challenge will be visualization, said Pak Chung Wong, chief scientist at the Department of Energy's Pacific Northwest National Laboratory. Visualization has long been a proven technique to help humans pinpoint trends and unusual events buried in large amounts of data, such as log files.

Standard visualization techniques may not work well with petabyte and exabyte-sized datasets, Wong warned. Such datasets may be arranged in hierarchies that can go 60 levels deep. "How can you represent that?" he asked.

Joab Jackson covers enterprise software and general technology breaking news for The IDG News Service. Follow Joab on Twitter at @Joab_Jackson. Joab's e-mail address is Joab_Jackson@idg.com

---

Agenda for the NIST Data Science Symposium 2014

National Institute of Standards and Technology
100 Bureau Drive, Gaithersburg, MD 20899
Building 101, Red Auditorium

PDF (Revised for Two Hour Delay)
Day 1 - March 4, 2014

Arrival
8:00 Check-in (Administration Building 101)
Poster-setup (Hall of Flags)

Welcome & Opening Remarks
John Garofolo - Senior Advisor for Program Development, IAD (NIST)
8:45-9:00 Charles Romine - Director of the Information Technology Laboratory (NIST)

The NIST Data Science Program and Symposium Objectives
9:00-9:15 Ashit Talukder - Chief, Information Access Division (NIST)
Overview of the NIST Data Science Program

9:15-9:30 Craig Greenberg - Project Leader, Multimodal Information Group (NIST)
Overview of the NIST Data Science Evaluation and Metrology Plans

9:30-9:40 A Discussion

Challenges and Gaps in Data Science Technologies
9:40-9:45 Session Introduction

9:45-10:00 Doug Cutting - Chief Architect (Cloudera)
Topic: Scalable Computing

10:00-10:15 Haesun Park - Professor & Executive Director, Center for Data Analytics (Georgia Tech)
Data and Visual Analytics for Large Scale High Dimensional Data

10:15-10:30 Eric Frost - Director, Viz Center (San Diego State)
Topic: Large-Scale Complex Visualizations

10:30-10:45 Pak Chung Wong – Chief Scientist (Pacific Northwest National Laboratory)
Topic: Challenges in Very Large-Scale Visualization and Human-Computer Interaction

10:45-11:00 Q&A Discussion

11:00-11:25 Break and Networking
Data and Use Cases for Data Science Research

11:25 Session Introduction

11:30-11:45 Jeanne Holm - Evangelist (Data.gov)
Datasets that Enable Rigorous Data Science Research

11:45-12:00 Stanley Ahalt - Director, Renaissance Computing Institute (University of North Carolina)
Research Data Sets and Infrastructure: The Challenge of Taming Research Data for Better Science

12:00-12:15 Thomas Karl - Director, National Climatic Data Center (NOAA)
Climate Archives in NOAA: Challenges and Opportunities

12:15-12:30 Chelsey Richards - Deputy Director for Public Health Scientific Services (CDC)
Topic: Health Data, Challenges and Opportunities

12:30-12:45 A Discussion

12:45-John Garofolo - Working Lunch Logistics

Lunch

12:45-2:00 Lunch: NIST cafeteria accepts cash or credit
- Lecture Room B: Focus Group on HCI and Data Science
- Lecture Room E: Focus Group on Big Data Analytics for Smart Manufacturing Systems
- Portrait Room: Focus Group for NIST Data Science Program plans

Data Science Poster Session

2:00-3:25 Poster Presentations - Hall of Flags

Data Science Benchmarking and Measurement

3:25-3:30 Session Introduction

3:30-3:45 Milind Bhandarkar - Chief Scientist (Pivotal Software Inc.)
Deep Analytics Pipeline: A Benchmark Proposal

3:45 4:00 Shahram Ghandeharizadeh - Associate Professor & Director Database Laboratory (University of South California)
Topic: Data Benchmarking

4:00-4:15 Shashi Shekhar - Professor (University of Minnesota)
Topic: Spatial Data Mining and Spatial Data Benchmarking
4:15-4:30 Volker Markl - Professor and Chair of The Database Systems & Information Management Group (TU Berlin)
Data Science Research from the Database Management Systems Perspective

4:30-4:45 Michael Hurley - Technical Staff (MIT Lincoln Laboratory)
Information Theoretic Evaluation of Data Processing Systems

4:45-5:00 Q&A Discussion

5:00 John Garofolo - Brief wrap-up and adjourn

---

**Day 2 - March 5, 2014**

**Arrival**

8:00-9:00(Administration Building 101)

**Keynote Address**

Charles Romine - Director of the Information Technology Laboratory (NIST)
Welcome and Introductions

9:00-9:30 Dr. Francine Berman - Chair, Research Data Alliance / U.S. Edward P. Hamilton Distinguished Professor in Computer Science
Director, Center for a Digital Society, Rensselaer Polytechnic Institute

9:30-9:45 Q&A Discussion

**Breakout Sessions**

9:45 Mark Przybocki - Group Leader, Multimodal Information Group (NIST)
Breakout Instructions

10:00-11:30 Breakouts

Data Science Benchmarking & Performance Measurement
Craig Greenberg (Chair)
Portrait Room

Datasets and Use Cases for Data Science Research
Ashit Talukder (Chair)
Green Auditorium

Challenges and Gaps in Data Science Technologies
John Garofolo (Chair)
Red Auditorium

11:30 John Garofolo - Working Lunch Logistics

Lunch

11:30-1:00 Lunch and Focused Topic Working Lunches
Lecture Room E: Focus Group on Uncertainty Quantification for Experimental Data

Government Data Science R&D Panel

1:00-2:45 Christopher White - Program Manager, XDATA Program (DARPA)
- Jason Matheny - Program Manager, Open Source Indicators (OSI) Program (IARPA)
- Kenneth Rice – Chief, ISR Integration Division (NGA)
- Lucy Nowell - Program Manager, Advanced Scientific Computing Research (DOE)
- Mike Huerta - Associate Director, US National Library of Medicine (NIH)
- Stephen Dennis- Innovation Director (HSARPA)
- Xiaoming Huo - Program Director (NSF)

2:45-3:00 Q&A Discussion

3:00-3:15 Break

Breakout Reports

3:15-3:25 Data Science Benchmarking & Performance Measurement
3:25-3:35 Datasets and Use Cases for Data Science Research
3:35-3:45 Challenges and Gaps in Data Science Technologies
3:45-3:55 Focus Group: Uncertainty Quantification for Experimental Data (Anne Plant)
3:55-4:05 Focus Group: HCI and Data Science (Sharon Laskowski)
4:05-4:15 Focus Group: Big Data Analytics for Smart Manufacturing Systems (Sudarsan Rachuri)

4:15-4:30 Discussion & Next Steps; Facilitator: Ashit Talukder

Abstract

Word

What a Data Scientist Does and How They Do It

Brand L. Niemann, Semantic Community
I am going to tell and show What a Data Scientist Does and How They Do It with the following topics:

- The State of Federal Data Science
- Data Science Team Examples
- Data Science, Data Products, and Data Stories Venn Diagrams
- NAS Report on Frontiers in Massive Data Analysis
- Graph Databases and the Semantic Web
- Semantic Medline – YarcData Graph Appliance Application for Federal Big Data Senior Steering WG
- Federal Big Data Working Group Meetup

I am both a data scientist and a data journalist. I engage in a wide range of data science activities, organize and participate in Data Science Teams and Meetups, and use data science to produce data products and stories like this one for the NIST Data Science Symposium. I will also show highlights of four recent Blog Posts to Data Science DC and Semantic Community.

The Semantic Medline – YarcData Graph Appliance Application for the Federal Big Data Senior Steering WG topics includes:

- Graphs and Traditional Technologies and The YarcData Approach
- Semantic Medline at NIH-NLM and Bioinformatics Publication
- Semantic Medline Database Application and Work Flow
- Predication Structure for Text Extraction of Triples
- Visualization and Linking to Original Text
- New Use Cases and Making the Most of Big Data

Two YouTube Videos will be shown (Schizo-7 minutes, Cancer-21 minutes).

Finally some highlights of the new Federal Big Data Working Group Meetups will be shown. All are welcome to participate in this new Meetup and their upcoming workshop in June.

The Fourth Meetup will be held tonight (march 4th) at The National Science Foundation, 4201 Wilson Boulevard, Arlington, Virginia 22230, from 6:30-9:30 p.m. NIH Welcome and Introduction to Biomedical Big Data Research: NIH Program Director by Dr. Peter Lyster. Brief Demo of NIH Semantic Medline/YarcData by Drs. Tom Rindflesch and Ted Slater. NSF Welcome and Introduction by NITRD Program Office Director, Dr. George Strawn. Presentation on BRAIN by Dr. Barend Mons. Discussion of A Data Fairport Workshop Summary. Open Discussion. Networking. RSVP: bniemann@cox.net

Bio

Brand Niemann, former Senior Enterprise Architect & Data Scientist with the US EPA, works as a data scientist, produces data science products, publishes data stories for Semantic Community, AOL Government, & Data Science & Data Visualization DC, and co-organizes the new Federal Big Data Working Group Meetup with Kate Goodier.
What a Data Scientist Does and How They Do It

Abstract

I am going to tell and show you What a Data Scientist Does and How They Do It with the following topics:

• The State of Federal Data Science
• Data Science Team Examples
• Data Science, Data Products, and Data Stories Venn Diagrams
• NAS Report on Frontiers in Massive Data Analysis
• Graph Databases and the Semantic Web
• Semantic Medline – YarcData Graph Appliance Application for Federal Big Data Senior Steering WG

My brief Bio is found below.

I am both a data scientist and a data journalist. I engage in a wide range of data science activities (Slides 4-6), organize and participate in Data Science Teams (Slides 7-8), and use data science to produce data products and stories (Slides 9-16) like this one for the NIST Data Science Symposium, November 18-19, 2013. I show highlights of four recent Blog Posts to Data Science DC and Semantic Community.

In my most recent Blog Post, I have suggested a mapping between the three Venn Diagrams for describing Data Science as shown in the table below.

<table>
<thead>
<tr>
<th>Drew Conway-Data Science</th>
<th>Harlan Harris-Data Products</th>
<th>Josh Tauberer &amp; Brand Niemann-Data Stories</th>
<th>Recent Conferences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Substantive Expertise</td>
<td>Domain Knowledge</td>
<td>Find and Prepare Data Sets</td>
<td>Data Transparency 2013</td>
</tr>
<tr>
<td>Hacking Skills</td>
<td>Software Engineering</td>
<td>Store and Query Data Sets</td>
<td>* Cloud: SOA, Semantics, and Data Science Conference</td>
</tr>
<tr>
<td>Math and Statistics Knowledge</td>
<td>Statistics, Predictive Analytics, and Visualization</td>
<td>Discover Data Stories in the Data Sets</td>
<td>Predictive Analytics World 2013 Government Conference</td>
</tr>
</tbody>
</table>

* Tried to cover all three.

These are really just three different ways of describing something that is an evolving discipline and science influenced by the personal experience of three data scientists with different experience.

I participate in the NIST Big Data Public Working Group (NBD-PWG), have suggested they follow the recent National Academy of Sciences (NAS) Report on Frontiers in Massive Data Analysis, July, 2013 (PDF and Wiki versions) because...
it takes a data science point-of-view, is written by expert data scientists, and provides the same items that the NBD-PWG needs (Slides 17-24):

- Definition
- Taxonomy
- Reference Architecture
- Technical Architecture
- Roadmap

My data story for the NBD-PWG provides a Roadmap (Slide 23) and concludes they need to Make a Technology Roadmap Project a Data Project (Slide 24).

The Graph Databases and the Semantic Web topic includes (Slides 25-30):

- New Book and Tutorial on Neo4j and My Talking Points
- Knowledge Base of NIST Symposium and BD-PWG Use Cases
- Semantic Web Linked Data for Triples
- Spotfire Cover Page and NIST Big Data Public Work Group Use Cases and Knowledge Bases

Essentially, I did a Relational and Graph Database pilot using the NIST Big Data Public Work Group Use Cases and Knowledge Bases. Big Data Uses Cases need to include Graph Databases since they contain stronger semantic relationships than RDBMS, reduce or eliminate the problems with RDMS, and because most Hadoop-type applications need them as well.

I found only one Use Case that uses Graph Databases: 34 Enabling Face-Book like Semantic Graph-search on Scientific Chemical and Text-based Data; Talapady Bhat, NIST (Summary and Full).

Finally, the Semantic Medline – YarcData Graph Appliance Application for Federal Big Data Senior Steering WG topic includes (Slides 31-40):

- Graphs and Traditional Technologies and The YarcData Approach
- Semantic Medline at NIH-NLM and Bioinformatics Publication
- Semantic Medline Database Application and Work Flow
- Predication Structure for Text Extraction of Triples
- Visualization and Linking to Original Text
- New Use Cases and Making the Most of Big Data

A YouTube Video of the Live Demo given at the recent Cloud: SOA, Semantics, and Data Science Conference is in process. The slides are available (Tim White and Tom Rindflesch). Update: YarcData Videos (Schizophrenia-7 minutes, Cancer-21 minutes)

Bio

Brand Niemann, former Senior Enterprise Architect and Data Scientist with the US EPA, completed 30 years of federal service in 2010. Since then he has worked as a data scientist for a number of organizations, produced data science
products for a large number of data sets, and published data stories for Federal Computer Week, Semantic Community and AOL Government.

Slides

Slide 1 What a Data Scientist Does and How They Do It

http://semanticommunity.info/
http://breakinggov.com/author/brand-niemann/
http://semanticommunity.info/Data_Science/Data_Science_Symposium_2013#Story

What a Data Scientist Does and How They Do It

Dr. Brand Niemann
Director and Senior Data Scientist
Semantic Community
http://semanticommunity.info/
AOL Government Brains
http://breakinggov.com/author/brand-niemann/
November 18, 2013

http://semanticommunity.info/Data_Science/Data_Science_Symposium_2013#Story

Slide 2 Overview

Overview

- I am Going to Tell and Show You What a Data Scientist Does and How They Do It With the Following Topics:
  - The State of Federal Data Science
  - Data Science Team Examples
  - Data Science, Data Products, and Data Stories Venn Diagrams
  - NAS Report on Frontiers in Massive Data Analysis
  - Graph Databases and the Semantic Web
  - Semantic Medline – YarcData Graph Appliance Application for Federal Big Data Senior Steering WG
Slide 3 Brief Bio

**Brief Bio**

- Former Senior Enterprise Architect and Data Scientist with the US EPA. Completed 30 years of federal service in 2010.
- Since then worked as a data scientist for a number of organizations, produced data science products for a large number of data sets, and published data stories for Federal Computer Week, Semantic Community and AOL Government.

---

Slide 4 The State of Federal Data Science: My Activities

**The State of Federal Data Science:**

**My Activities**

- **Data Science Products and Stories:**
  - Big data and data science conferences and employment opportunities.
- **Data Science Journalism:**
  - Federal Computer Week, Government Computer News, AOL Government News (now BreakingGov), and Information Week.
- **Data Science for Governments:**
- **Data Science for the White House and OMB:**
  - Project Open Government Data and Tool Requirements.
- **Federal Big Data Senior Steering Work Group:**
  - A Semantic Web Strategy for Big Data: Semantic Medline on the Yarc Graph Computer (January and upcoming).
- **W3C eGovernance Data Science Community:**
  - Proposed this recently.
- **Data Science DC and Graph Database Meetups:**
  - Attend meetings and write event reviews and data product stories.

---

Slide 5 The State of Federal Data Science: What is Big Data?

http://semanticommunity.info/
The State of Federal Data Science: What is Big Data?

- Dominic Sale, new OMB Chief of Data Analytics & Reporting, said the new Digital Government Strategy is “treating all content as data.”
- So big data = all your content and here is my matrix for the Third Annual Government Big Data Forum.

Source: [http://semanticommunity.info/](http://semanticommunity.info/)

---

Slide 6 The State of Federal Data Science: My Big Data Matrix

[http://semanticommunity.info/Third_Annual_Government_Big_Data_Forum](http://semanticommunity.info/Third_Annual_Government_Big_Data_Forum)

The State of Federal Data Science: My Big Data Matrix

```
<table>
<thead>
<tr>
<th>Four Vs</th>
<th>Concept</th>
<th>Method</th>
<th>Goal</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volume and Velocity</td>
<td>Big Data=All Content</td>
<td>Make all content as data</td>
<td>Federal Digital Government Strategy</td>
<td>Knowledge Base</td>
</tr>
<tr>
<td>Veracity</td>
<td>Web-Linked Data</td>
<td>Semantic Web Data</td>
<td>Strong Relationships</td>
<td>Spreadsheets &amp; Relational and Graph Databases</td>
</tr>
<tr>
<td>Value</td>
<td>Unified Data Architecture</td>
<td>Data Integration</td>
<td>Data Ecosystem Network</td>
<td>Visualizations</td>
</tr>
</tbody>
</table>
```

Source: [http://semanticommunity.info/Third_Annual_Government_Big_Data_Forum](http://semanticommunity.info/Third_Annual_Government_Big_Data_Forum)

---

Slide 7 Data Science Team Example: Open Government Data Manager

[http://semanticommunity.info/An_Open_Data_Policy#Story](http://semanticommunity.info/An_Open_Data_Policy#Story)
Data Science Team Example: Chief Data Science Officer

Chief Data Science Officer:
- Dr. George Strawn, Director, White House OSTP  
  NITRD/NCO: Semantic Medline could be the "killer" Semantic Web application for the US Federal Government

Data Science Team:
- Dr. Brand Niemann, Lead  
- Dr. Tom Rindflesch, NLM Semantic Medline Creator  
- Professor Kirk Borne, George Mason University  
  • Federal Big Data Senior Steering WG Workforce Training Initiative  
- Tim White, Director, YarcData Federal Global Head  
- Aaron Bossett, YarcData Federal Solution Architect  
- Dr. Eric Little, Modus Operandi Chief Scientist

http://semanticommunity.info/A_NITRD_Dashboard/Making_the_Most_of_Big_Data#Story
Slide 9 Data Science, Data Products, and Data Stories Venn Diagrams

Data Science, Data Products, and Data Stories Venn Diagrams

- Data Science DC Community Posts:
  - Cloud SOA Semantics and Data Science Conference
    - Semantic Medline – YarcData Graph Appliance Pilot to discover new relationships between disease and treatments for mental illness, cancer, etc.
  - Confidential Data Event Review
  - Data Science DC In Review
    - All of their content=big data
  - Predictive Analytics Government Conference
    - Data Science, Data Products, and Predictive Analytics

Slide 10 Data Science DC Community Posts: Cloud SOA Semantics and Data Science Conference

http://datacommunitydc.org/blog/2013...ce-conference/

Slide 11 Data Science DC Community Posts: Event Review: Confidential Data

http://datacommunitydc.org/blog/2013...idential-data/
Slide 12 Data Science DC Community Posts: Data Science DC In Review

http://semanticommunity.info/Data_Science/Data_Science_DC#Story_2

Slide 13 Data Science DC Community Posts: Data Science DC Data in Spotfire Web Player

Web Player
Slide 14 Data Science DC Community Posts: Data Science, Data Products, and Predictive Analytics

http://semanticommunity.info/Analytics/Predictive_Analytic_World_Government_2013#Story
Slide 15 Data Science DC Community Posts: Data Science, Data Products, and Predictive Analytics 2

Data Science DC Community Posts:
Data Science, Data Products, and Predictive Analytics

The mapping between the three Venn Diagrams is shown in the table below:

<table>
<thead>
<tr>
<th>Drew Conway - Data Science</th>
<th>Harlan Harris - Data Products</th>
<th>Josh Tauberer &amp; Brand Niemann - Data Stories</th>
<th>Recent Conferences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Substantive Expertise</td>
<td>Domain Knowledge</td>
<td>Find and Prepare Data Sets</td>
<td>Data Transparency 2013</td>
</tr>
<tr>
<td>Hacking Skills</td>
<td>Software Engineering</td>
<td>Store and Query Data Sets</td>
<td>Cloud: SOA, Semantics, and Data Science Conference</td>
</tr>
<tr>
<td>Math and Statistics</td>
<td>Statistics, Predictive Analytics, and Visualization</td>
<td>Discover Data Stories in the Data Sets</td>
<td>Predictive Analytics World 2013 Government Conference</td>
</tr>
</tbody>
</table>

* Tried to cover all three. These are really just three different ways of describing something that is an evolving discipline and science influenced by the personal experience of three data scientists with different experience.

Slide 16 Data Science DC Community Posts: Data Science, Data Products, and Predictive Analytics 3

Web Player

Slide 17 NAS Report on Frontiers in Massive Data Analysis

http://semanticommunity.info/Big_Data_at_NIST/Frontiers_in_Massive_Data_Analysis

http://semanticommunity.info/Data_Science/Data_Science_Symposium_2013
Updated: Wed, 23 Sep 2015 02:59:49 GMT
Powered by mindtouch™
NAS Report on Frontiers in Massive Data Analysis

- The reason I like this recent Report is that it takes the data scientists point of view (written by some real expert data scientists) and reflects how they work:
  - Application to the NIST Big Data Public Working Group
  - Definition
  - Taxonomy
  - Reference Architecture
  - Technical Architecture
  - Roadmap
  - Make a Technology Roadmap Project a Data Project
NAS Report on Frontiers in Massive Data Analysis: Definition

- This report uses the terms "big data" and "massive data" interchangeably to refer to data at massive scale.
  - Source: [Footnote 1]
- Indeed, massive data sets generally involve growth not merely in the number of individuals represented (the "rows" of the database) but also in the number of descriptors of those individuals (the "columns" of the database). Moreover, we are often interested in the predictive ability associated with combinations of the descriptors; this can lead to exponential growth in the number of hypotheses considered, with severe consequences for error rates. That is, a naive appeal to a "law of large numbers" for massive data is unlikely to be justified; if anything, the perils associated with statistical fluctuations may actually increase as data sets grow in size.
  - Source: [The Promise and Perils of Massive Data. Also see: Big Data Buzzwords from & to For specific examples see: Table 1.1 Scientific and Engineering Fields Impacted by Massive Data and My Cross-Walk Table (Agency Managers with Big Problems and Big Sets)]
- My Comment: Simply put: conventional analytics, statistics, and visualizations may not work.

Slide 20 NAS Report on Frontiers in Massive Data Analysis: Taxonomy

NAS Report on Frontiers in Massive Data Analysis: Taxonomy

- "As part of the study that led to this report, the Committee on the Analysis of Massive Data developed a taxonomy of some of the major algorithmic problems arising in massive data analysis.
  - 1. Basic statistics,
  - 2. Generalized N-body problem,
  - 3. Graph-theoretic computations,
  - 4. Linear algebraic computations,
  - 5. Optimization,
  - 6. Integration, and
  - 7. Alignment problems."

Source: [Conclusions]

Slide 21 NAS Report on Frontiers in Massive Data Analysis: Reference Architecture
NAS Report on Frontiers in Massive Data Analysis: Reference Architecture

- Use their topics and content:
  - Scaling the Infrastructure for Data Management
  - Temporal Data and Real-time Algorithms
  - Large-Scale Data Representations
  - Resources, Trade-offs, and Limitations
  - Building Models from Massive Data
  - Sampling and Massive Data
  - Human Interaction with Data
  - The Seven Computational Giants of Massive Data Analysis

Source: URLs above

---

NAS Report on Frontiers in Massive Data Analysis: Technical Architecture

- "For each of these computational classes, there are computational constraints that arise within any particular problem domain that help to determine the specialized algorithmic strategy to be employed. Most work in the past has focused on a setting that involves a single processor with the entire data set fitting in random access memory (RAM). Additional important settings for which algorithms are needed include the following:
  - The streaming setting, in which data arrive in quick succession, and only a subset can be stored;
  - The disk-based setting, in which the data are too large to store in RAM but fit on one machine's disk;
  - The distributed setting, in which the data are distributed over multiple machines' RAMs or disks; and
  - The multi-threaded setting, in which the data lie on one machine having multiple processors that share RAM."

Source: Conclusions
Slide 23 NAS Report on Frontiers in Massive Data Analysis: Roadmap

http://semanticommunity.info/Data_Science/Graph_Databases

NAS Report on Frontiers in Massive Data Analysis: Roadmap

- The Open Government Data Manager and/or Chief Data Science Officer gives the Data Science Team a business or scientific problem to solve by asking questions of the data.
- The Data Scientist Team decides:
  - Which of the 7 major algorithmic problems (Taxonomy) they are trying to solve to get the answer;
  - Which of the 8 major computational classes they need to address in handling the data (Reference Architecture); and
  - Which of the 4 major computational constraints (Technical Architecture) they need to address with the hardware.
- They also need to include Graph Databases since they contain stronger semantic relationships than RDBMS, reduce or eliminate the problems with RDBMS, and because most Hadoop-type applications need them as well.
- See Graph Databases Book and Tutorial

Slide 24 NAS Report on Frontiers in Massive Data Analysis: Make a Technology Roadmap Project a Data Project

http://semanticommunity.info/Big_Data_at_NIST#Story

Slide 25 Graph Databases and the Semantic Web: New Book and Tutorial on Neo4j

http://semanticommunity.info/Data_Science/Graph_Databases#Story
http://semanticommunity.info/Data_Science/Graph_Databases/Tutorial
Graph Databases and the Semantic Web: New Book and Tutorial on Neo4j

http://semanticommunity.info/Data_Science/Graph_Databases/Tutorial#Neo4j_Case_Studies
http://www.meetup.com/graphdb-baltim...nts/137240202/
http://semanticommunity.info/Data_Science/Graph_Databases#W3C_Support

Slide 26 Graph Databases and the Semantic Web: My Talking Points

http://semanticommunity.info/Data_Science/Graph_Databases/Tutorial#Neo4j_Case_Studies
http://www.meetup.com/graphdb-baltim...nts/137240202/
http://semanticommunity.info/Data_Science/Graph_Databases#W3C_Support

Graph Databases and the Semantic Web: My Talking Points

- Charts are not Graphs!
- Leonhard Euler 1707-1783 provided the mathematical theory
- SQL and RDBMS require multiple key tables to cross-walk for data integration.
  Graph databases do not.
- Start with a whiteboard. Add relationships. Add properties of the relationships. Model incrementally. Code it. Query it (Cypher) and visualize it in Neo4j.
- It is all about patterns. The Architecture is the Model is the Working Application!
- Neo4j is open source, free Java-based software
- See impressive Neo4j Case Studies
- Attend the next Graph Database Meetup, October 24th:
  - http://www.meetup.com/graphdb-baltimore/events/137240202/
- Statements in the book’s Appendix A. NOSQL about Graph Databases and the W3C’s Semantic Web led me to ask the questions:
  - Is Neo4j the best for Semantic Web Applications? and
  - Is Spotfire Network Analytics like Chapter 7 Predictive Analysis with Graph Theory?

Slide 27 Graph Databases and the Semantic Web: Knowledge Base of NIST Symposium and BD-PWG Use Cases

http://semanticommunity.info/Data_Science/Data_Science_Symposium_2013
http://semanticommunity.info/Data_Science/Data_Science_Symposium_2013/Full_Uses_Cases
Slide 28 Graph Databases and the Semantic Web: Semantic Web Linked Data for Triples

http://semanticommunity.info/api/deki/files/26427/NISTBigDataScience.xlsx

Slide 29 Graph Databases and the Semantic Web: Spotfire Cover Page

Web Player
Slide 30 Graph Databases and the Semantic Web: NIST Big Data Public Work
Group Use Cases and Knowledge Bases in Spotfire

Web Player
Slide 31 Semantic Medline – YarcData Graph Appliance Application for Federal Big Data Senior Steering WG: Graphs and Traditional Technologies

**Semantic Medline – YarcData Graph Appliance Application for Federal Big Data Senior Steering WG: Graphs and Traditional Technologies**

- **Square peg, round hole:**
  - Current technology does not support efficient representation, storage, and interaction with complex graph structures
  - Traditional relational models only add to the already complex structure
  - Traditional hardware approaches do not support efficient access to highly-interconnected graphs

- **You don’t know what you don’t know:**
  - Efficient relational schemas require prior knowledge of the relationships between database fields
  - Updating and modifying schemas frequently introduces delays and errors

- **Problems in partitioning the problem:**
  - Distributed computing solutions are good... if your problem can be easily partitioned
  - Graphs are not predictable; accessing graph nodes across large clusters can be unworkable at best and does not work at scale

---

Slide 32 Semantic Medline – YarcData Graph Appliance Application for Federal Big Data Senior Steering WG: The YarcData Approach

**Semantic Medline – YarcData Graph Appliance Application for Federal Big Data Senior Steering WG: The YarcData Approach**

**Business Challenge:**

- Large Shared Memory Architecture
  - Up to 512 TB
- XMTZ Massively Multi-Threaded Processors
  - 128 Threads
- Scalable I/O
  - Up to 350TB per Hour

Real-time, Interactive Analytics on Large Graph Problems
Slide 33 Semantic Medline – YarcData Graph Appliance Application for Federal Big Data Senior Steering WG: Semantic Medline at NIH-NLM

Semantic Medline – YarcData Graph Appliance Application for Federal Big Data Senior Steering WG: Semantic Medline at NIH-NLM

- Current: Web based research tool.
- Transition: Current systems re-engineered to leverage Indra (less than 5 days)
- Purpose: Build a platform users to perform increasingly complex analysis
- Immediate Requirement: Replicate current capability
- Future: Allow for increasingly complex analysis. Ability to capture and share analytics in addition to sharing data. Tailor Indra to less complex queries.

Slide 34 Semantic Medline – YarcData Graph Appliance Application for Federal Big Data Senior Steering WG: Bioinformatics Publication


Slide 35 Semantic Medline – YarcData Graph Appliance Application for Federal Big Data Senior Steering WG: Semantic Medline Database Application

Slide 36 Semantic Medline – YarcData Graph Appliance Application for Federal Big Data Senior Steering WG: Work Flow

Semantic Medline – YarcData Graph Appliance Application for Federal Big Data Senior Steering WG: Work Flow

Document retrieval
MEDLINE citations
SemRep
Semantic predications
Automatic summarization
Graphical summary
Biomedical information management
Slide 37: Semantic Medline – YarcData Graph Appliance Application for Federal Big Data Senior Steering WG: Predication Structure for Text Extraction of Triples

Slide 38: Semantic Medline – YarcData Graph Appliance Application for Federal Big Data Senior Steering WG: Visualization and Linking to Original Text
Slide 39: Semantic Medline – YarcData Graph Appliance Application for Federal Big Data Senior Steering WG: New Use Cases

- Schizophrenia
  - Current therapies target dopamine receptors
    - Not entirely effective
    - Side effects
  - Basic research is exploring glutamate and its NMDA receptor
  - Goal: can we use Semantic MEDLINE to discover that research trend in the scientific literature

- Cancer
  - With some exceptions, therapy is not effective
    - Has not progressed significantly in 50 years
  - Scientific basis
    - Traditionally — cancer cells
  - Immune system and cancer
    - Connection noted in 1869 (Vinckho)
    - But not exploited until recently
  - Goal: look for trends in cancer immunotherapy

Note: YouTube Video of Demo in Process

Slide 40: Semantic Medline – YarcData Graph Appliance Application for Federal Big Data Senior Steering WG: Making the Most of Big Data

http://semanticommunity.info/A_NITRD_Dashboard/Making_the_Most_of_Big_Data

Slide 41: Extra Slides: Galls’ Law


Gall's Law

http://semanticommunity.info/Data_Science/Data_Science_Symposium_2013
Updated: Wed, 23 Sep 2015 02:59:49 GMT
Powered by mindtouch®
Extra Slides: Gall’s Law

- So next time you hear a politician or senior government manager describe something really complicated they want to do, and for us to pay for, ask them do you have a simple dashboard for that to convince us they can make that work.

- There is a famous quote from John Gall known as Gall’s Law: “A complex system that works is invariably found to have evolved from a simple system that worked. A complex system designed from scratch never works and cannot be patched up to make it work. You have to start over, beginning with a working simple system.”

Slide 42 The Value Proposition of Agile Analytics
Slide 43 System of Systems Architecture

System of Systems Architecture
Dynamic Case Management (e.g. Be Informed)

Data Science Library (e.g. Spotfire)

Data Science Products (e.g. Spotfire)

Semantic Index of Linked Data (e.g. Excel)

Slide 44 Extra Slides: NGA Demo

http://semanticommunity.info/Network_Centricity/NGA_Demo#Story
http://semanticommunity.info/Be_Informed_Be_Free#Story

Extra Slides: NGA Demo

Spotfire Dashboard

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

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

Monday, September 30, 2013, NIST Big Data Workshop, Comments and Suggestions

http://semanticommunity.info/Big_Data_at_NIST

Data Science, Data Products, and Data Stories Venn Diagrams: http://semanticommunity.info/Analytics/Predictive_Analytic_World_Government_2013#Story

The reason I like the recent NAS Report on Frontiers in Massive Data Analysis is that it takes the data scientists point of view (written by some real expert data scientists) and reflects how they work:

First, Scientific and Engineering Fields Impacted by Big Data: See Noteworthy Use Cases (like Professor Fox’s Use Case but more generic)

http://semanticommunity.info/Big_Data_at_NIST/Frontiers_in_Massive_Data_Analysis#TABLE_1.1_Scientific_and_Engineering_Fields_Impacted_by_Massive_Data

Second, the need for someone like a Chief Data Officer to provide the question(s) they want the Data Science Team and big database (s) to answer:
http://semanticommunity.info/A_NITRD_Dashboard/Making_the_Most_of_Big_Data

Our Semantic Medline – YARCDATA Graph Appliance Application for Dr. George Strawn’s Federal Big Data Senior Steering WG to discover new relationships between disease and treatments for mental illness, cancer, etc.
http://semanticommunity.info/Data_Science/Cloud_SOA_Semantics_and_Data_Science_Conference (see slides at end of first day – video of demo in process)

Third, explain that “massive data sets generally involve growth not merely in the number of individuals represented (the “rows” of the database) but also in the number of descriptors of those individuals (the “columns” of the database). Moreover, we are often interested in the predictive ability associated with combinations of the descriptors; this can lead to exponential growth in the number of hypotheses considered, with severe consequences for error rates. That is, a naive appeal to a “law of large numbers” for massive data is unlikely to be justified; if anything, the perils associated with statistical fluctuations may actually increase as data sets grow in size.” Simply put: conventional analytics, statistics, and visualizations may not work.

Fourth, follow the taxonomy, reference architecture, and technical architecture outlines:
http://semanticommunity.info/Big_Data_at_NIST/Frontiers_in_Massive_Data_Analysis#Story

so the Data Scientists decides which of the 7 major algorithmic problems (Taxonomy) they are trying to solve to get the answer;

which of the 8 major computational classes they need to address in handling the data (Reference Architecture); and

which of the 4 major computational constraints (Technical Architecture) they need to address with the hardware.
Also the need to include Graph Databases since they contain stronger semantic relationships than RDBMS, reduce or eliminate the problems with RDMS, and because most Hadoop-type applications need them as well: http://semanticommunity.info/Data_Science/Graph_Databases

---

**Announcement of Data Science Symposium 2013**


**Summary:**

Given the explosion of data production, storage capabilities, communications technologies, computational power, and supporting infrastructure, data science is now recognized as a highly-critical growth area with impact across many sectors including science, government, finance, health care, manufacturing, advertising, retail, and others. Since data science technologies are being leveraged to drive crucial decision making, it is of paramount importance to be able to measure the performance of these technologies and to correctly interpret their output. The NIST Information Technology Laboratory is forming a cross-cutting data science program focused on driving advancements in data science through system benchmarking and rigorous measurement science.

**Description:**

The inaugural NIST Data Science Symposium will convene a diverse multi-disciplinary community of stakeholders to promote the design, development, and adoption of novel measurement science in order to foster advances in Big Data processing, analytics, visualization, interaction, and lifecycle management. It is set apart from related symposia by our emphasis on advancing data science technologies through:

- Benchmarking of complex data-intensive analytic systems and subcomponents
- Developing general, extensible performance metrics and measurement methods
- Creating reference datasets & challenge problems grounded in rigorous measurement science
- Coordination of open, community-driven evaluations that focus on domains of general interest.

**Location:**

The first symposium will be held November 18-19, 2013 on the NIST campus in Gaithersburg, MD.

**Organizing Committee:**

Ashit Talukder (NIST), John Garofolo (NIST), Mark Przybocki, (NIST), Craig Greenberg (NIST)

**Registration:**

Registration to attend the NIST Data Science Symposium is now open. Registration is free, but it is necessary to register in order to attend. **The deadline for registration will be on or before Monday, November 11.** Registration may close once the capacity of the venue is reached. Please note that only registered participants will be permitted to enter the NIST campus to attend the symposium. To register, please go to: [http://semanticommunity.info/Data_Science/Data_Science_Symposium_2013](http://semanticommunity.info/Data_Science/Data_Science_Symposium_2013)
Call For Abstracts:

Participants who wish to give presentations of their technical perspectives or present posters (potentially with technical demonstrations) that address symposium topics should submit a brief one-page abstract and brief one-paragraph bio to datascience@nist.gov by October 4th, 2013. Submitters will be notified whether their perspectives have been selected for plenary or poster presentation by October 18th. Speakers, panelists, and poster presenters will be selected by the organizers based on relevance to symposium objectives and workshop balance. Due to the technical nature of the symposium, no marketing will be permitted.

Symposium Topics:

Understanding the Data Science Technical Landscape:

• Primary challenges in and technical approaches to complex workflow components of Big Data systems, including ETL, lifecycle management, analytics, visualization & human-system interaction.
• Major forms of analytics employed in data science.

Improving Analytic System Performance via Measurement Science

• Generation of ground truth for large datasets and performance measurement with limited or no ground truth.
• Methods to measure the performance of data analytic workflows where there are multiple subcomponents, decision points, and human interactions.
• Methods to measure the flow of uncertainty across complex data analytic systems.
• Approaches to formally characterizing end-to-end analytic workflows.

Datasets to Enable Rigorous Data Science Research

• Useful properties for data science reference datasets.
• Leveraging simulated data in data science research.
• Efficient approaches to sharing research data.

Contact

NIST maintains a general mailing list for our Data Science Measurement and Evaluation program: datascience-list@nist.gov

Relevant information is posted to this list. If you would like to join the list or have any question for NIST related to our data science program, please email us at: datascience@nist.gov
1. Introduction

2. Use Case Summaries

Government Operation

2.1 Census 2010 and 2000 – Title 13 Big Data; Vivek Navale & Quyen Nguyen, NARA

Application: Preserve Census 2010 and 2000 – Title 13 data for a long term in order to provide access and perform analytics after 75 years. One must maintain data “as-is” with no access and no data analytics for 75 years; one must preserve the data at the bit-level; one must perform curation, which includes format transformation if necessary; one must provide access and analytics after nearly 75 years. Title 13 of U.S. code authorizes the Census Bureau and guarantees that individual and industry specific data is protected.

Current Approach: 380 terabytes of scanned documents

2.2 National Archives and Records Administration Accession NARA, Search, Retrieve, Preservation; Vivek Navale & Quyen Nguyen, NARA


Current Approach: 1) Get physical and legal custody of the data; 2) Pre-process data for virus scan, identifying file format identification, removing empty files; 3) Index; 4) Categorize records (sensitive, unsensitive, privacy data, etc.); 5) Transform old file formats to modern formats (e.g. WordPerfect to PDF); 6) E-discovery; 7) Search and retrieve to respond to special request; 8) Search and retrieve of public records by public users. Currently 100’s of terabytes stored centrally in commercial databases supported by custom software and commercial search products.

Futures: There are distributed data sources from federal agencies where current solution requires transfer of those data to a centralized storage. In the future, those data sources may reside in multiple Cloud environments. In this case, physical custody should avoid transferring big data from Cloud to Cloud or from Cloud to Data Center.
2.3 Statistical Survey Response Improvement (Adaptive Design); Cavan Capps, U.S. Census Bureau

Application: Survey costs are increasing as survey response declines. The goal of this work is to use advanced “recommendation system techniques” that are open and scientifically objective, using data mashed up from several sources and historical survey para-data (administrative data about the survey) to drive operational processes in an effort to increase quality and reduce the cost of field surveys.

Current Approach: About a petabyte of data coming from surveys and other government administrative sources. Data can be streamed with approximately 150 million records transmitted as field data streamed continuously, during the decennial census. All data must be both confidential and secure. All processes must be auditable for security and confidentiality as required by various legal statutes. Data quality should be high and statistically checked for accuracy and reliability throughout the collection process. Use Hadoop, Spark, Hive, R, SAS, Mahout, Allegrograph, MySQL, Oracle, Storm, BigMemory, Cassandra, Pig software.

Futures: Need to improve recommendation systems similar to those used in e-commerce (see Netflix use case) that reduce costs and improve quality while providing confidentiality safeguards that are reliable and publically auditable. Data visualization is useful for data review, operational activity and general analysis. It continues to evolve; mobile access important.

2.4 Non-Traditional Data in Statistical Survey Response Improvement (Adaptive Design); Cavan Capps, U.S. Census Bureau

Application: Survey costs are increasing as survey response declines. This use case has similar goals to that above but involves non-traditional commercial and public data sources from the web, wireless communication, electronic transactions mashed up analytically with traditional surveys to improve statistics for small area geographies, new measures and to improve the timeliness of released statistics.

Current Approach: Integrate survey data, other government administrative data, web scrapped data, wireless data, e-transaction data, potentially social media data and positioning data from various sources. Software, Visualization and data characteristics similar to previous use case.

Futures: Analytics needs to be developed which give statistical estimations that provide more detail, on a more near real time basis for less cost. The reliability of estimated statistics from such “mashed up” sources still must be evaluated.

Commercial

2.5 Cloud Eco-System, for Financial Industries (Banking, Securities & Investments, Insurance) transacting business within the United States; Pw Carey, Compliance Partners, LLC

Application: Use of Cloud (Bigdata) technologies needs to be extended in Financial Industries (Banking, Securities & Investments, Insurance).

Current Approach: Currently within Financial Industry, Bigdata and Hadoop are used for fraud detection, risk analysis and assessments as well as improving the organizations knowledge and understanding of their customers. At the same time, the traditional client/server/data warehouse/RDBM (Relational Database Management) systems are used for the
handling, processing, storage and archiving of the entities financial data. Real time data and analysis is important in these applications.

**Futures:** One must address GRC (Governance, Risk & Compliance) as well as CIA (Confidentiality, Integrity & Availability) issues, which can even be impacted by the SEC’s mandated use of XBRL (extensible Business Related Markup Language). While at the same time, addressing the influence these same issues can and will have within a Cloud Eco-system/Big Data environment, and their global impact across all Financial sectors.

### 2.6 Mendeley – An International Network of Research; William Gunn, Mendeley

**Application:** Mendeley has built a database of research documents and facilitates the creation of shared bibliographies. Mendeley uses the information collected about research reading patterns and other activities conducted via the software to build more efficient literature discovery and analysis tools. Text mining and classification systems enables automatic recommendation of relevant research, improving the cost and performance of research teams, particularly those engaged in curation of literature on a particular subject.

**Current Approach:** Data size is 15TB presently, growing about 1 TB/month. Processing on Amazon Web Services with Hadoop, Scribe, Hive, Mahout, Python. Standard libraries for machine learning and analytics, Latent Dirichlet Allocation, custom built reporting tools for aggregating readership and social activities per document.

**Futures:** Currently Hadoop batch jobs are scheduled daily, but work has begun on real-time recommendation. The database contains ~400M documents, roughly 80M unique documents, and receives 5-700k new uploads on a weekday. Thus a major challenge is clustering matching documents together in a computationally efficient way (scalable and parallelized) when they’re uploaded from different sources and have been slightly modified via third-part annotation tools or publisher watermarks and cover pages.

### 2.7 Netflix Movie Service; Geoffrey Fox, Indiana University

**Application:** Allow streaming of user selected movies to satisfy multiple objectives (for different stakeholders) -- especially retaining subscribers. Find best possible ordering of a set of videos for a user (household) within a given context in real-time; maximize movie consumption. Digital movies stored in cloud with metadata; user profiles and rankings for small fraction of movies for each user. Use multiple criteria – content based recommender system; user-based recommender system; diversity. Refine algorithms continuously with A/B testing.

**Current Approach:** Recommender systems and streaming video delivery are core Netflix technologies. Recommender systems are always personalized and use logistic/linear regression, elastic nets, matrix factorization, clustering, latent Dirichlet allocation, association rules, gradient boosted decision trees and others. Winner of Netflix competition (to improve ratings by 10%) combined over 100 different algorithms. Uses SQL, NoSQL, MapReduce on Amazon Web Services. Netflix recommender systems have features in common to e-commerce like Amazon. Streaming video has features in common with other content providing services like iTunes, Google Play, Pandora and Last.fm.

**Futures:** Very competitive business. Need to aware of other companies and trends in both content (which Movies are hot) and technology. Need to investigate new business initiatives such as Netflix sponsored content.
2.8 Web Search; Geoffrey Fox, Indiana University

**Application:** Return in ~0.1 seconds, the results of a search based on average of 3 words; important to maximize quantities like "precision@10" or number of great responses in top 10 ranked results.

**Current Approach:** Steps include 1) Crawl the web; 2) Pre-process data to get searchable things (words, positions); 3) Form Inverted Index mapping words to documents; 4) Rank relevance of documents: PageRank; 5) Lots of technology for advertising, "reverse engineering ranking" "preventing reverse engineering"; 6) Clustering of documents into topics (as in Google News) 7) Update results efficiently. Modern clouds and technologies like MapReduce have been heavily influenced by this application. ~45B web pages total.

**Futures:** A very competitive field where continuous innovation needed. Two important areas are addressing mobile clients which are a growing fraction of users and increasing sophistication of responses and layout to maximize total benefit of clients, advertisers and Search Company. The "deep web" (that behind user interfaces to databases etc.) and multimedia search of increasing importance. 500M photos uploaded each day and 100 hours of video uploaded to YouTube each minute.

2.9 IaaS (Infrastructure as a Service) Big Data Business Continuity & Disaster Recovery (BC/DR) Within a Cloud Eco-System; Pw Carey, Compliance Partners, LLC

**Application:** BC/DR (Business Continuity/Disaster Recovery) needs to consider the role that the following four overlaying and inter-dependent forces will play in ensuring a workable solution to an entity’s business continuity plan and requisite disaster recovery strategy. The four areas are; people (resources), processes (time/cost/ROI), technology (various operating systems, platforms and footprints) and governance (subject to various and multiple regulatory agencies).

**Current Approach:** Cloud Eco-systems, incorporating IaaS (Infrastructure as a Service), supported by Tier 3 Data Centers provide data replication services. Replication is different from Backup and only moves the changes since the last time a replication occurs, including block level changes. The replication can be done quickly, with a five second window, while the data is replicated every four hours. This data snap shot is retained for seven business days, or longer if necessary. Replicated data can be moved to a Fail-over Center to satisfy an organizations RPO (Recovery Point Objectives) and RTO (Recovery Time Objectives). Technologies from VMware, NetApps, Oracle, IBM, and Brocade are some of those relevant. Data sizes are terabytes up to petabytes.

**Futures:** The complexities associated with migrating from a Primary Site to either a Replication Site or a Backup Site is not fully automated at this point in time. The goal is to enable the user to automatically initiate the Failover sequence. Both organizations must know which servers have to be restored and what are the dependencies and inter-dependencies between the Primary Site servers and Replication and/or Backup Site servers. This requires a continuous monitoring of both.

2.10 Cargo Shipping; William Miller, MaCT USA

**Application:** Monitoring and tracking of cargo as in Fedex, UPS and DHL.
Current Approach: Today the information is updated only when the items that were checked with a bar code scanner are sent to the central server. The location is not currently displayed in real-time. An architectural diagram is in full description.

Futures: This Internet of Things application needs to track items in real time. A new aspect will be status condition of the items which will include sensor information, GPS coordinates, and a unique identification schema based upon a new ISO 29161 standards under development within ISO JTC1 SC31 WG2.

Figure 0 Cargo Shipping

![Cargo Shipping Diagram](http://semanticommunity.info/Data_Science/Data_Science_Symposium_2013)

2.11 Materials Data for Manufacturing; John Rumble, R&R Data Services

Application: Every physical product is made from a material that has been selected for its properties, cost, and availability. This translates into hundreds of billion dollars of material decisions made every year. However the adoption of new materials normally takes decades (two to three) rather than a small number of years, in part because data on new materials is not easily available. One needs to broaden accessibility, quality, and usability and overcome proprietary barriers to sharing materials data. One must create sufficiently large repositories of materials data to support discovery.

Current Approach: Currently decisions about materials usage are unnecessarily conservative, often based on older rather than newer materials R&D data, and not taking advantage of advances in modeling and simulations.

Futures: Materials informatics is an area in which the new tools of data science can have major impact by predicting the performance of real materials (gram to ton quantities) starting at the atomistic, nanometer, and/or micrometer level of description. One must establish materials data repositories beyond the existing ones that focus on fundamental data; one must develop internationally-accepted data recording standards that can be used by a very diverse materials community, including developers materials test standards (such as ASTM and ISO), testing companies, materials producers, and R&D labs; one needs tools and procedures to help organizations wishing to deposit proprietary materials in data repositories to mask proprietary information, yet to maintain the usability of data; one needs multi-variable materials data visualization tools, in which the number of variables can be quite high.
2.12 Simulation driven Materials Genomics; David Skinner, LBNL

Application: Innovation of battery technologies through massive simulations spanning wide spaces of possible design. Systematic computational studies of innovation possibilities in photovoltaics. Rational design of materials based on search and simulation. These require management of simulation results contributing to the materials genome.

Current Approach: PyMatGen, FireWorks, VASP, ABINIT, NWChem, BerkeleyGW, and varied materials community codes running on large supercomputers such as 150K core Hopper machine at NERSC produce results that are not synthesized.

Futures: Need large scale computing at scale for simulation science. Flexible data methods at scale for messy data. Machine learning and knowledge systems that integrate data from publications, experiments, and simulations to advance goal-driven thinking in materials design. The current 100TB of data will become 500TB in 5 years.

Defense

2.13 Large Scale Geospatial Analysis and Visualization; David Boyd, Data Tactics

Application: Need to support large scale geospatial data analysis and visualization. As the number of geospatially aware sensors increase and the number of geospatially tagged data sources increases the volume geospatial data requiring complex analysis and visualization is growing exponentially.

Current Approach: Traditional GIS systems are generally capable of analyzing a millions of objects and easily visualizing thousands. Data types include Imagery (various formats such as NITF, GeoTiff, CADRG), and vector with various formats like shape files, KML, text streams. Object types include points, lines, areas, polylines, circles, ellipses. Data accuracy very important with image registration and sensor accuracy relevant. Analytics include closest point of approach, deviation from route, and point density over time, PCA and ICA. Software includes Server with Geospatially enabled RDBMS, Geospatial server/analysis software – ESRI ArcServer, GeoServer; Visualization by ArcMap or browser based visualization.

Futures: Today’s intelligence systems often contain trillions of geospatial objects and need to be able to visualize and interact with millions of objects. Critical issues are Indexing, retrieval and distributed analysis; Visualization generation and transmission; Visualization of data at the end of low bandwidth wireless connections; Data is sensitive and must be completely secure in transit and at rest (particularly on handhelds); Geospatial data requires unique approaches to indexing and distributed analysis.

2.14 Object identification and tracking from Wide Area Large Format Imagery (WALF) Imagery or Full Motion Video (FMV) – Persistent Surveillance; David Boyd, Data Tactics

Application: Persistent surveillance sensors can easily collect petabytes of imagery data in the space of a few hours. The data should be reduced to a set of geospatial object (points, tracks, etc.) which can easily be integrated with other data to form a common operational picture. Typical processing involves extracting and tracking entities (vehicles, people, packages) over time from the raw image data.
Current Approach: It is unfeasible for this data to be processed by humans for either alerting or tracking purposes. The data needs to be processed close to the sensor which is likely forward deployed since it is too large to be easily transmitted. Typical object extraction systems are currently small (1-20 node) GPU enhanced clusters. There are a wide range of custom software and tools including traditional RDBMS’s and display tools. Real time data obtained at FMV (Full Motion Video) – 30-60 frames per/sec at full color 1080P resolution or WALF (Wide Area Large Format) with 1-10 frames per/sec at 10Kx10K full color resolution. Visualization of extracted outputs will typically be as overlays on a geospatial (GIS) display. Analytics are basic object detection analytics and integration with sophisticated situation awareness tools with data fusion. Significant security issues so that sources and methods cannot be compromised so the enemy should not be able to know what we see

Futures: Typical problem is integration of this processing into a large (GPU) cluster capable of processing data from several sensors in parallel and in near real time. Transmission of data from sensor to system is also a major challenge.

2.15 Intelligence Data Processing and Analysis; David Boyd, Data Tactics

Application: Allow Intelligence Analysts to a) Identify relationships between entities (people, organizations, places, equipment) b) Spot trends in sentiment or intent for either general population or leadership group (state, non-state actors) c) Find location of and possibly timing of hostile actions (including implantation of IEDs) d) Track the location and actions of (potentially) hostile actors e) Ability to reason against and derive knowledge from diverse, disconnected, and frequently unstructured (e.g. text) data sources f) Ability to process data close to the point of collection and allow data to be shared easily to/from individual soldiers, forward deployed units, and senior leadership in garrison.

Current Approach: Software includes Hadoop, Accumulo (Big Table), Solr, Natural Language Processing, Puppet (for deployment and security) and Storm running on medium size clusters. Data size in 10s of Terabytes to 100s of Petabytes with Imagery intelligence device gathering petabyte in a few hours. Dismounted warfighters would have at most 1-100s of Gigabytes (typically handheld data storage).

Futures: Data currently exists in disparate silos which must be accessible through a semantically integrated data space. Wide variety of data types, sources, structures, and quality which will span domains and requires integrated search and reasoning. Most critical data is either unstructured or imagery/video which requires significant processing to extract entities and information. Network quality, Provenance and security essential.

Healthcare and Life Sciences

2.16 Electronic Medical Record (EMR) Data; Shaun Grannis, Indiana University

Application: Large national initiatives around health data are emerging, and include developing a digital learning health care system to support increasingly evidence-based clinical decisions with timely accurate and up-to-date patient-centered clinical information; using electronic observational clinical data to efficiently and rapidly translate scientific discoveries into effective clinical treatments; and electronically sharing integrated health data to improve healthcare process efficiency and outcomes. These key initiatives all rely on high-quality, large-scale, standardized and aggregate health data. One needs advanced methods for normalizing patient, provider, facility and clinical concept identification within and among separate health care organizations to enhance models for defining and extracting clinical phenotypes...
from non-standard discrete and free-text clinical data using feature selection, information retrieval and machine learning decision-models. One must leverage clinical phenotype data to support cohort selection, clinical outcomes research, and clinical decision support.

**Current Approach:** Clinical data from more than 1,100 discrete logical, operational healthcare sources in the Indiana Network for Patient Care (INPC) the nation's largest and longest-running health information exchange. This describes more than 12 million patients, more than 4 billion discrete clinical observations. > 20 TB raw data. Between 500,000 and 1.5 million new real-time clinical transactions added per day.

**Futures:** Teradata, PostgreSQL, MongoDB running on Indiana University supercomputer supporting information retrieval methods to identify relevant clinical features (tf-idf, latent semantic analysis, mutual information). Natural Language Processing techniques to extract relevant clinical features. Validated features will be used to parameterize clinical phenotype decision models based on maximum likelihood estimators and Bayesian networks. Decision models will be used to identify a variety of clinical phenotypes such as diabetes, congestive heart failure, and pancreatic cancer.

### 2.17 Pathology Imaging/digital pathology; Fusheng Wang, Emory University

**Application:** Digital pathology imaging is an emerging field where examination of high resolution images of tissue specimens enables novel and more effective ways for disease diagnosis. Pathology image analysis segments massive (millions per image) spatial objects such as nuclei and blood vessels, represented with their boundaries, along with many extracted image features from these objects. The derived information is used for many complex queries and analytics to support biomedical research and clinical diagnosis. The full description has examples of 2-D and 3-D pathology images.

**Current Approach:** 1GB raw image data + 1.5GB analytical results per 2D image. MPI for image analysis; MapReduce + Hive with spatial extension on supercomputers and clouds. GPU’s used effectively. The full description shows the architecture of Hadoop-GIS, a spatial data warehousing system over MapReduce to support spatial analytics for analytical pathology imaging.

**Futures:** Recently, 3D pathology imaging is made possible through 3D laser technologies or serially sectioning hundreds of tissue sections onto slides and scanning them into digital images. Segmenting 3D microanatomic objects from registered serial images could produce tens of millions of 3D objects from a single image. This provides a deep “map” of human tissues for next generation diagnosis. 1TB raw image data + 1TB analytical results per 3D image and 1PB data per moderated hospital per year.

**Figure 1:** Examples of 2-D and 3-D pathology images

![2D Digital Pathology](image1.png)

![3D Digital Pathology](image2.png)
2.18 Computational Bioimaging; David Skinner, Joaquin Correa, Daniela Ushizima, Joerg Meyer, LBNL

**Application:** Data delivered from bioimaging is increasingly automated, higher resolution, and multi-modal. This has created a data analysis bottleneck that, if resolved, can advance the biosciences discovery through Big Data techniques.

**Current Approach:** The current piecemeal analysis approach does not scale to situation where a single scan on emerging machines is 32TB and medical diagnostic imaging is annually around 70 PB excluding cardiology. One needs a web-based one-stop-shop for high performance, high throughput image processing for producers and consumers of models built on bio-imaging data.

**Futures:** Our goal is to solve that bottleneck with extreme scale computing with community-focused science gateways to support the application of massive data analysis toward massive imaging data sets. Workflow components include data acquisition, storage, enhancement, minimizing noise, segmentation of regions of interest, crowd-based selection and extraction of features, and object classification, and organization, and search. Use ImageJ, OMERO, VolRover, advanced segmentation and feature detection software.

2.19 Genomic Measurements; Justin Zook, NIST

**Application:** NIST/Genome in a Bottle Consortium integrates data from multiple sequencing technologies and methods to develop highly confident characterization of whole human genomes as reference materials, and develop methods to use these Reference Materials to assess performance of any genome sequencing run.

**Current Approach:** The storage of ~40TB NFS at NIST is full; there are also PBs of genomics data at NIH/NCBI. Use Open-source sequencing bioinformatics software from academic groups (UNIX-based) on a 72 core cluster at NIST supplemented by larger systems at collaborators.

**Futures:** DNA sequencers can generate ~300GB compressed data/day which volume has increased much faster than Moore’s Law. Future data could include other ‘omics’ measurements, which will be even larger than DNA sequencing. Clouds have been explored.
2.20 Comparative analysis for metagenomes and genomes; Ernest Szeto, LBNL (Joint Genome Institute)

**Application:** Given a metagenomic sample, (1) determine the community composition in terms of other reference
isolate genomes, (2) characterize the function of its genes, (3) begin to infer possible functional pathways, (4)
characterize similarity or dissimilarity with other metagenomic samples, (5) begin to characterize changes in community
composition and function due to changes in environmental pressures, (6) isolate sub-sections of data based on quality
measures and community composition.

**Current Approach:** integrated comparative analysis system for metagenomes and genomes, front ended by an
interactive Web UI with core data, backend precomputations, batch job computation submission from the UI. Provide
interface to standard bioinformatics tools (BLAST, HMMER, multiple alignment and phylogenetic tools, gene callers,
sequence feature predictors…).

**Futures:** Management of heterogeneity of biological data is currently performed by relational database management
system (Oracle). Unfortunately, it does not scale for even the current volume 50TB of data. NoSQL solutions aim at
providing an alternative but unfortunately they do not always lend themselves to real time interactive use, rapid and
parallel bulk loading, and sometimes have issues regarding robustness.

2.21 Individualized Diabetes Management; Ying Ding , Indiana University

**Application:** Diabetes is a growing illness in world population, affecting both developing and developed countries.
Current management strategies do not adequately take into account of individual patient profiles, such as co-morbidities
and medications, which are common in patients with chronic illnesses. Need to use advanced graph-based data mining
techniques applied to EHR converted into a RDF graph, to search for Diabetes patients and extract their EHR data for
outcome evaluation.

**Current Approach:** Typical patient data records composed of 100 controlled vocabulary values and 1000 continuous
values. Most values have a timestamp. Need to change traditional paradigm of relational row-column lookup to semantic
graph traversal.

**Futures:** Identify similar patients from a large Electronic Health Record (EHR) database, i.e. an individualized cohort,
and evaluate their respective management outcomes to formulate most appropriate solution suited for a given patient
with diabetes. Use efficient parallel retrieval algorithms, suitable for cloud or HPC, using open source Hbase with both
indexed and custom search to identify patients of possible interest. Use Semantic Linking for Property Values method to
convert an existing data warehouse at Mayo Clinic, called the Enterprise Data Trust (EDT), into RDF triples that enables
one to find similar patients through linking of both vocabulary-based and continuous values. The time dependent
properties need to be processed before query to allow matching based on derivatives and other derived properties.

2.22 Statistical Relational Artificial Intelligence for Health Care; Sriraam Natarajan, Indiana University

**Application:** The goal of the project is to analyze large, multi-modal medical data including different data types such as
imaging, EHR, genetic and natural language. This approach employs the relational probabilistic models that have the
capability of handling rich relational data and modeling uncertainty using probability theory. The software learns models
from multiple data types and can possibly integrate the information and reason about complex queries. Users can
provide a set of descriptions – say for instance, MRI images and demographic data about a particular subject. They can then query for the onset of a particular disease (say Alzheimer’s) and the system will then provide a probability distribution over the possible occurrence of this disease.

**Current Approach:** A single server can handle a test cohort of a few hundred patients with associated data of 100’s of GB.

**Futures:** A cohort of millions of patient can involve petabyte datasets. Issues include availability of too much data (as images, genetic sequences etc) that can make the analysis complicated. A major challenge lies in aligning the data and merging from multiple sources in a form that can be made useful for a combined analysis. Another issue is that sometimes, large amount of data is available about a single subject but the number of subjects themselves is not very high (i.e., data imbalance). This can result in learning algorithms picking up random correlations between the multiple data types as important features in analysis.

2.23 World Population Scale Epidemiological Study; Madhav Marathe, Stephen Eubank or Chris Barrett, Virginia Tech

**Application:** One needs reliable real-time prediction and control of pandemic similar to the 2009 H1N1 influenza. In general one is addressing contagion diffusion of various kinds: information, diseases, social unrest can be modeled and computed. All of them can be addressed by agent-based models that utilize the underlying interaction network to study the evolution of the desired phenomena.

**Current Approach:** (a) Build a synthetic global population. (b) Run simulations over the global population to reason about outbreaks and various intervention strategies. Current 100TB dataset generated centrally with MPI based simulation system written in Charm++. Parallelism is achieved by exploiting the disease residence time period.

**Futures:** Use large social contagion models to study complex global scale issues

2.24 Social Contagion Modeling for Planning, Public Health and Disaster Management; Madhav Marathe or Chris Kuhlman, Virginia Tech

**Application:** Model Social behavior including national security, public health, viral marketing, city planning, disaster preparedness. In a social unrest application, people take to the streets to voice unhappiness with government leadership. There are citizens that both support and oppose government. Quantify the degrees to which normal business and activities are disrupted owing to fear and anger. Quantify the possibility of peaceful demonstrations, violent protests. Quantify the potential for government responses ranging from appeasement, to allowing protests, to issuing threats against protestors, to actions to thwart protests. To address these issues, must have fine-resolution models (at level of individual people, vehicles, and buildings) and datasets.

**Current Approach:** The social contagion model infrastructure includes different types of human-to-human interactions (e.g., face-to-face versus online media) to be simulated. It takes not only human-to-human interactions into account, but also interactions among people, services (e.g., transportation), and infrastructure (e.g., internet, electric power). These activity models are generated from averages like census data.
Futures: Data fusion a big issue; how should one combine data from different sources and how to deal with missing or incomplete data? take into account heterogeneous features of 100s of millions or billions of individuals, models of cultural variations across countries that are assigned to individual agents? How to validate these large models?

2.25 Biodiversity and LifeWatch; Wouter Los, Yuri Demchenko, University of Amsterdam

Application: Research and monitor different ecosystems, biological species, their dynamics and migration with a mix of custom sensors and data access/processing and a federation with relevant projects in area. Particular case studies: Monitoring alien species, monitoring migrating birds, wetlands. See ENVRI for integration of LifeWatch with other environmental e-infrastructures.

Futures: LifeWatch initiative will provide integrated access to a variety of data, analytical and modeling tools as served by a variety of collaborating initiatives. Another service is offered with data and tools in selected workflows for specific scientific communities. In addition, LifeWatch will provide opportunities to construct personalized ‘virtual labs’, also allowing one to enter new data and analytical tools. New data will be shared with the data facilities cooperating with LifeWatch. LifeWatch operates the Global Biodiversity Information facility and Biodiversity Catalogue that is Biodiversity Science Web Services Catalogue. Data includes ‘omics, species information, ecological information (such as biomass, population density etc.), ecosystem data (such as CO2 fluxes. Algal blooming, water and soil characteristics)

Deep Learning and Social Media

2.26 Large-scale Deep Learning; Adam Coates, Stanford University

Application: Increase the size of datasets and models that can be tackled with deep learning algorithms. Large models (e.g., neural networks with more neurons and connections) combined with large datasets are increasingly the top performers in benchmark tasks for vision, speech, and Natural Language Processing. One needs to train a deep neural network from a large (>>1TB) corpus of data (typically imagery, video, audio, or text). Such training procedures often require customization of the neural network architecture, learning criteria, and dataset pre-processing. In addition to the computational expense demanded by the learning algorithms, the need for rapid prototyping and ease of development is extremely high.

Current Approach: The largest applications so far are to image recognition and scientific studies of unsupervised learning with 10 million images and up to 11 billion parameters on a 64 GPU HPC Infiniband cluster. Both supervised (using existing classified images) and unsupervised applications investigated.

Futures: Large datasets of 100TB or more may be necessary in order to exploit the representational power of the larger models. Training a self-driving car could take 100 million images at megapixel resolution. Deep Learning shares many characteristics with the broader field of machine learning. The paramount requirements are high computational throughput for mostly dense linear algebra operations, and extremely high productivity for researcher exploration. One needs integration of high performance libraries with high level (python) prototyping environments.
2.27 Organizing large-scale, unstructured collections of consumer photos; David Crandall, Indiana University

Application: Produce 3D reconstructions of scenes using collections of millions to billions of consumer images, where neither the scene structure nor the camera positions are known a priori. Use resulting 3D models to allow efficient and effective browsing of large-scale photo collections by geographic position. Use 3D models to perform object recognition on each image. 3D reconstruction can be posed as a robust non-linear least squares optimization problem in which observed (noisy) correspondences between images are constraints and unknowns are 6-d camera pose of each image and 3-d position of each point in the scene.

Current Approach: Hadoop cluster with 480 cores processing data of initial applications. Note over 500 billion images on Facebook and over 5 billion on Flickr with over 500 million images added to social media sites each day.

Futures: Need many analytics including feature extraction, feature matching, and large-scale probabilistic inference, which appear in many or most computer vision and image processing problems, including recognition, stereo resolution, and image denoising. Need to visualize large-scale 3D reconstructions, and navigate large-scale collections of images that have been aligned to maps.

2.28 Truthy: Information diffusion research from Twitter Data; Filippo Menczer, Alessandro Flammini, Emilio Ferrara, Indiana University

Application: Understanding how communication spreads on socio-technical networks. Detecting potentially harmful information spread at the early stage (e.g., deceiving messages, orchestrated campaigns, untrustworthy information, etc.)

Current Approach: 1) Acquisition and storage of a large volume (30 TB a year compressed) of continuous streaming data from Twitter (~100 million messages per day, ~500GB data/day increasing over time); (2) near real-time analysis of such data, for anomaly detection, stream clustering, signal classification and online-learning; (3) data retrieval, big data visualization, data-interactive Web interfaces, public API for data querying. Use Python/SciPy/NumPy/MPI for data analysis. Information diffusion, clustering, and dynamic network visualization capabilities already exist.

Futures: Truthy plans to expand incorporating Google+ and Facebook. Need to move towards Hadoop/IndexedHBase & HDFS distributed storage. Use Redis as an in-memory database as a buffer for real-time analysis. Need streaming clustering, anomaly detection and online learning.

2.29 Crowd Sourcing in the Humanities as Source for Big and Dynamic Data; Sebastian Drude, Max-Planck-Institute for Psycholinguistics, Nijmegen The Netherlands

Application: Capture information (manually entered, recorded multimedia, reaction times, pictures, sensor information) from many individuals and their devices and so characterize wide ranging individual, social, cultural and linguistic variation among several dimensions (space, social space, time).

Current Approach: Use typically XML technology, traditional relational databases, and besides pictures not much multimedia yet.
Futures: Crowd sourcing has been barely started to be used on a larger scale but with the availability of mobile devices, now there is a huge potential for collecting much data from many individuals, also making use of sensors in mobile devices. This has not been explored on a large scale so far; existing projects of crowd sourcing are usually of a limited scale and web-based. Privacy issues may be involved (A/V from individuals), anonymization may be necessary but not always possible. Data management and curation critical. Size could be hundreds of terabytes with multimedia.

2.30 CINET: Cyberinfrastructure for Network (Graph) Science and Analytics; Madhav Marathe or Keith Bisset, Virginia Tech

Application: CINET provides a common web-based platform for accessing various (i) network and graph analysis tools such as SNAP, NetworkX, Galib, etc. (ii) real-world and synthetic networks, (iii) computing resources and (iv) data management systems to the end-user in a seamless manner.

Current Approach: CINET uses an Infiniband connected high performance computing cluster with 720 cores to provide HPC as a service. It is being used for research and education.

Futures: As the repository grows, we expect a rapid growth to lead to over 1000-5000 networks and methods in about a year. As more fields use graphs of increasing size, parallel algorithms will be important. Data manipulation and bookkeeping of the derived data for users is a challenge there are no well-defined and effective models and tools for management of various graph data in a unified fashion

2.31 NIST Information Access Division analytic technology performance measurement, evaluations, and standards; John Garofolo, NIST

Application: Develop performance metrics, measurement methods, and community evaluations to ground and accelerate the development of advanced analytic technologies in the areas of speech and language processing, video and multimedia processing, biometric image processing, and heterogeneous data processing as well as the interaction of analytics with users. Typically employ one of two processing models: 1) Push test data out to test participants and analyze the output of participant systems, 2) Push algorithm test harness interfaces out to participants and bring in their algorithms and test them on internal computing clusters.

Current Approach: Large annotated corpora of unstructured/semi-structured text, audio, video, images, multimedia, and heterogeneous collections of the above including ground truth annotations for training, developmental testing, and summative evaluations. The test corpora exceed 900M Web pages occupying 30 TB of storage, 100M tweets, 100M ground-truthed biometric images, several hundred thousand partially ground-truthed video clips, and terabytes of smaller fully ground-truthed test collections.

Futures: Even larger data collections are being planned for future evaluations of analytics involving multiple data streams and very heterogeneous data. As well as larger datasets, future includes testing of streaming algorithms with multiple heterogeneous data. Use of clouds being explored.

The Ecosystem for Research
2.32 DataNet Federation Consortium DFC; Reagan Moore, University of North Carolina at Chapel Hill

Application: Promote collaborative and interdisciplinary research through federation of data management systems across federal repositories, national academic research initiatives, institutional repositories, and international collaborations. The collaboration environment runs at scale: petabytes of data, hundreds of millions of files, hundreds of millions of metadata attributes, tens of thousands of users, and a thousand storage resources.

Current Approach: Currently 25 science and engineering domains have projects that rely on the iRODS (Integrated Rule Oriented Data System) policy-based data management system including major NSF projects such as Ocean Observatories Initiative (sensor archiving); Temporal Dynamics of Learning Center (Cognitive science data grid); the iPlant Collaborative (plant genomics); Drexel engineering digital library; Odum Institute for social science research (data grid federation with Dataverse). iRODS currently manages petabytes of data, hundreds of millions of files, hundreds of millions of metadata attributes, tens of thousands of users, and a thousand storage resources. It interoperates with workflow systems (NCSA Cyberintegrator, Kepler, Taverna), cloud and more traditional storage models and different transport protocols. The full description has a diagram of the iRODS architecture.

Figure Policy-based Data Management Concept Graph (iRODS)

2.33 The ‘Discinnet process’, metadata <-> big data global experiment; P. Journeau, Discinnet Labs

Application: Discinnet has developed a web 2.0 collaborative platform and research prototype as a pilot installation now becoming deployed to be appropriated and tested by researchers from a growing number and diversity of research fields through communities belonging a diversity of domains.

Its goal is to reach a wide enough sample of active research fields represented as clusters—researchers projected and aggregating within a manifold of mostly shared experimental dimensions—to test general, hence potentially interdisciplinary, epistemological models throughout the present decade.
Current Approach: Currently 35 clusters started with close to 100 awaiting more resources and potentially much more open for creation, administration and animation by research communities. Examples range from optics, cosmology, materials, microalgae, health to applied math, computation, rubber and other chemical products/issues.

Futures: Discinnet itself would not be Bigdata but rather will generate metadata when applied to a cluster that involves Bigdata. In interdisciplinary integration of several fields, the process would reconcile metadata from many complexity levels.

2.34 Enabling Face-Book like Semantic Graph-search on Scientific Chemical and Text-based Data; Talapady Bhat, NIST

Application: Establish social media-based infrastructure, terminology and semantic data-graphs to annotate and present technology information using ‘root’ and rule-based methods used primarily by some Indo-European languages like Sanskrit and Latin.

Current approach: Many reports, including a recent one on Material Genome Project finds that exclusive top-down solutions to facilitate data sharing and integration are not desirable for federated multi-disciplinary efforts. However, a bottom-up approach can be chaotic. For this reason, there is need for a balanced blend of the two approaches to support easy-to-use techniques to metadata creation, integration and sharing. This challenge is very similar to the challenge faced by language developer at the beginning. One of the successful effort used by many prominent languages is that of ‘roots’ and rules that form the framework for creating on-demand words for communication. In this approach a top-down method is used to establish a limited number of highly re-usable words called ‘roots’ by surveying the existing best practices in building terminology. These ‘roots’ are combined using few ‘rules’ to create terms on-demand by a bottom-up step.

Y(uj) (join), O (creator, God, brain), Ga (motion, initiation) –leads to ‘Yoga’ in Sanskrit, English

Geno (genos)-cide–race based killing – Latin, English

Bio-technology –English, Latin


A press release by the American Institute of Physics on this approach is at http://www.eurekalert.org/pub_releases/2013-07/aiop-ffm071813.php

Our efforts to develop automated and rule and root-based methods (Chem-BLAST - http://pdb.nist.gov/chemblast/pdb.pl) to identify and use best-practice, discriminating terms in generating semantic data-graphs for science started almost a decade back with a chemical structure database. This database has millions of structures obtained from the Protein Data Bank and the PubChem used world-wide. Subsequently we extended our efforts to build root-based terms to text-based data of cell-images. In this work we use few simple rules to define and extend terms based on best-practice as decided by weaning through millions of popular use-cases chosen from over hundred biological ontologies.

Currently we are working on extending this method to publications of interest to Material Genome, Open-Gov and NIST-wide publication archive - NIKE. - http://pdb.nist.gov/nike/term.pl. These efforts are a component of Research Data
Futures: Create a cloud infrastructure for social media of scientific information where many scientists from various parts of the world can participate and deposit results of their experiment. Some of the issues that one has to resolve prior to establishing a scientific social media are: a) How to minimize challenges related to establishing re-usable, inter-disciplinary, scalable, on-demand, use-case and user-friendly vocabulary? b) How to adopt a existing or create new on-demand ‘data-graph’ to place an information in an intuitive way such that it would easily integrate with existing ‘data-graphs’ in a federated environment without knowing too much about the data management? c) How to find relevant scientific data without spending too much time on the internet? Start with resources like the Open Government movement, Material genome Initiative and Protein Databank. This effort includes many local and networked resources. Developing an infrastructure to automatically integrate information from all these resources using data-graphs is a challenge that we are trying to solve. Good database tools and servers for data-graph manipulation are needed.

2.35 Light source beamlines; Eli Dart, LBNL

Application: Samples are exposed to X-rays from light sources in a variety of configurations depending on the experiment. Detectors (essentially high-speed digital cameras) collect the data. The data are then analyzed to reconstruct a view of the sample or process being studied.

Current Approach: A variety of commercial and open source software is used for data analysis – examples including Octopus for Tomographic Reconstruction, Avizo (http://vsq3d.com) and FIJI (a distribution of ImageJ) for Visualization and Analysis. Data transfer is accomplished using physical transport of portable media (severely limits performance) or using high-performance GridFTP, managed by Globus Online or workflow systems such as SPADE.

Futures: Camera resolution is continually increasing. Data transfer to large-scale computing facilities is becoming necessary because of the computational power required to conduct the analysis on time scales useful to the experiment. Large number of beamlines (e.g. 39 at LBNL ALS) means that aggregate data load is likely to increase significantly over the coming years and need for a generalized infrastructure for analyzing gigabytes per second of data from many beamline detectors at multiple facilities.

Astronomy and Physics

2.36 Catalina Real-Time Transient Survey (CRTS): a digital, panoramic, synoptic sky survey; S. G. Djorgovski, Caltech

Application: The survey explores the variable universe in the visible light regime, on time scales ranging from minutes to years, by searching for variable and transient sources. It discovers a broad variety of astrophysical objects and phenomena, including various types of cosmic explosions (e.g., Supernovae), variable stars, phenomena associated with accretion to massive black holes (active galactic nuclei) and their relativistic jets, high proper motion stars, etc. The data are collected from 3 telescopes (2 in Arizona and 1 in Australia), with additional ones expected in the near future (in Chile).
Current Approach: The survey generates up to ~0.1 TB on a clear night with a total of ~100 TB in current data holdings. The data are preprocessed at the telescope, and transferred to Univ. of Arizona and Caltech, for further analysis, distribution, and archiving. The data are processed in real time, and detected transient events are published electronically through a variety of dissemination mechanisms, with no proprietary withholding period (CRTS has a completely open data policy). Further data analysis includes classification of the detected transient events, additional observations using other telescopes, scientific interpretation, and publishing. In this process, it makes a heavy use of the archival data (several PB’s) from a wide variety of geographically distributed resources connected through the Virtual Observatory (VO) framework.

Futures: CRTS is a scientific and methodological testbed and precursor of larger surveys to come, notably the Large Synoptic Survey Telescope (LSST), expected to operate in 2020’s and selected as the highest-priority ground-based instrument in the 2010 Astronomy and Astrophysics Decadal Survey. LSST will gather about 30 TB per night. The schematic architecture for a cyber-infrastructure for time domain astronomy illustrated by a figure in full description.

Figure Catalina Real-Time Transient Survey (CRTS): a digital, panoramic, synoptic sky survey

Figure: One possible schematic architecture for a cyber-infrastructure for time domain astronomy. Transient event data streams are produced by survey pipelines from the telescopes on the ground or in space, and the events with their observational descriptions are ingested by one or more depositories, from which they can be disseminated electronically to human astronomers or robotic telescopes. Each event is assigned an evolving portfolio of information, which would include all of the available data on that celestial position, from a wide variety of data archives unidied under the Virtual Observatory framework, expert annotations, etc. Representations of such federated information can be both human-readable and machine-readable. They are fed into one or more automated event characterization, classification, and prioritization engines that deploy a variety of machine learning tools for these tasks. Their output, which evolves dynamically as new information arrives and is processed, informs the follow-up observations of the selected events, and the resulting data are communicated back to the event portfolios, for the next iteration. Users (human or robotic) can tap into the system at multiple points, both for an information retrieval, and to contribute new information, through a standardized set of formats and protocols. This could be done in a (near) real time, or in an archival (not time critical) modes.

2.37 DOE Extreme Data from Cosmological Sky Survey and Simulations; Salman Habib, Argonne National Laboratory; Andrew Connolly, University of Washington

Application: A cosmology discovery tool that integrates simulations and observation to clarify the nature of dark matter, dark energy, and inflation, some of the most exciting, perplexing, and challenging questions facing modern physics
including the properties of fundamental particles affecting the early universe. The simulations will generate comparable data sizes to observation.

**Futures:** Data sizes are Dark Energy Survey (DES) 4 PB in 2015; Zwicky Transient Factory (ZTF) 1 PB/year in 2015; Large Synoptic Sky Survey (LSST see CRTS description) 7 PB/year in 2019; Simulations > 10 PB in 2017. Huge amounts of supercomputer time (over 200M hours) will be used.

2.38 Large Survey Data for Cosmology; Peter Nugent LBNL

**Application:** For DES (Dark Energy Survey) the data are sent from the mountaintop via a microwave link to La Serena, Chile. From there, an optical link forwards them to the NCSA as well as NERSC for storage and "reduction". Here galaxies and stars in both the individual and stacked images are identified, catalogued, and finally their properties measured and stored in a database.

**Current Approach:** Subtraction pipelines are run using extant imaging data to find new optical transients through machine learning algorithms. Linux cluster, Oracle RDBMS server, Postgres PSQL, large memory machines, standard Linux interactive hosts, GPFS. For simulations, HPC resources. Standard astrophysics reduction software as well as Perl/Python wrapper scripts, Linux Cluster scheduling.

**Futures:** Techniques for handling Cholesky decomposition for thousands of simulations with matrices of order 1M on a side and parallel image storage would be important. LSST will generate 60PB of imaging data and 15PB of catalog data and a correspondingly large (or larger) amount of simulation data. Over 20TB of data per night.

2.39 Particle Physics: Analysis of LHC Large Hadron Collider Data: Discovery of Higgs particle; Michael Ernst BNL, Lothar Bauerdick FNAL, Geoffrey Fox, Indiana University; Eli Dart, LBNL

**Application:** Study CERN LHC (Large Hadron Collider) Accelerator and Monte Carlo producing events describing particle-apparatus interaction. Processed information defines physics properties of events (lists of particles with type and momenta). These events are analyzed to find new effects; both new particles (Higgs) and present evidence that conjectured particles (Supersymmetry) have not been detected. LHC has a few major experiments including ATLAS and CMS. These experiments have global participants (for example CMS has 3600 participants from 183 institutions in 38 countries), and so the data at all levels is transported and accessed across continents.

**Current Approach:** The LHC experiments are pioneers of a distributed Big Data science infrastructure, and several aspects of the LHC experiments' workflow highlight issues that other disciplines will need to solve. These include automation of data distribution, high performance data transfer, and large-scale high-throughput computing. Grid analysis with 350,000 cores running "continuously" over 2 million jobs per day arranged in 3 tiers (CERN, "Continents/ Countries". “Universities”). Uses "Distributed High Throughput Computing" (Pleasing parallel) architecture with facilities integrated across the world by WLCG (LHC Computing Grid) and Open Science Grid in the US. 15 Petabytes data gathered each year from Accelerator data and Analysis with 200PB total. Specifically in 2012 ATLAS had at Brookhaven National Laboratory (BNL) 8PB Tier1 tape; BNL over 10PB Tier1 disk and US Tier2 centers 12PB disk cache. CMS has similar data sizes. Note over half resources used for Monte Carlo simulations as opposed to data analysis.
Futures: In the past the particle physics community has been able to rely on industry to deliver exponential increases in performance per unit cost over time, as described by Moore’s Law. However the available performance will be much more difficult to exploit in the future since technology limitations, in particular regarding power consumption, have led to profound changes in the architecture of modern CPU chips. In the past software could run unchanged on successive processor generations and achieve performance gains that follow Moore’s Law thanks to the regular increase in clock rate that continued until 2006. The era of scaling HEP sequential applications is now over. Changes in CPU architectures imply significantly more software parallelism as well as exploitation of specialized floating point capabilities. The structure and performance of HEP data processing software needs to be changed such that it can continue to be adapted and further developed in order to run efficiently on new hardware. This represents a major paradigm-shift in HEP software design and implies large scale re-engineering of data structures and algorithms. Parallelism needs to be added at all levels at the same time, the event level, the algorithm level, and the sub-algorithm level. Components at all levels in the software stack need to interoperate and therefore the goal is to standardize as much as possible on basic design patterns and on the choice of a concurrency model. This will also help to ensure efficient and balanced use of resources.

Figure 1 The LHC Collider location at CERN

![Figure 1 The LHC Collider location at CERN](image)

Figure 2 The Multi-tier LHC computing infrastructure

![Figure 2 The Multi-tier LHC computing infrastructure](image)
2.40 Belle II High Energy Physics Experiment; David Asner & Malachi Schram, PNNL

Application The Belle experiment is a particle physics experiment with more than 400 physicists and engineers investigating CP-violation effects with B meson production at the High Energy Accelerator KEKB e+ e- accelerator in Tsukuba, Japan. In particular look at numerous decay modes at the Upsilon(4S) resonance to search for new phenomena beyond the Standard Model of Particle Physics. This accelerator has the largest intensity of any in the world but events simpler than those from LHC and so analysis is less complicated but similar in style compared to the CERN accelerator.

Futures: An upgraded experiment Belle II and accelerator SuperKEKB will start operation in 2015 with a factor of 50 increased data with total integrated RAW data ~120PB and physics data ~15PB and ~100PB MC samples. Move to a distributed computing model requiring continuous RAW data transfer of ~20Gbps at designed luminosity between Japan and US. Will need Open Science Grid, Geant4, DIRAC, FTS, Belle II framework software.

Earth, Environmental and Polar Science

2.41 EISCAT 3D incoherent scatter radar system; Yin Chen, Cardiff University; Ingemar Häggström, Ingrid Mann, Craig Heinselman, EISCAT Science Association

Application: EISCAT, the European Incoherent Scatter Scientific Association, conducts research on the lower, middle and upper atmosphere and ionosphere using the incoherent scatter radar technique. This technique is the most powerful ground-based tool for these research applications. EISCAT studies instabilities in the ionosphere, as well as investigating the structure and dynamics of the middle atmosphere. It is also a diagnostic instrument in ionospheric modification experiments with addition of a separate Heating facility. Currently EISCAT operates 3 of the 10 major incoherent radar scattering instruments worldwide with its facilities in in the Scandinavian sector, north of the Arctic Circle.

Current Approach: The current running old EISCAT radar generates terabytes per year rates and does not present special challenges.

Futures: The design of the next generation radar, EISCAT_3D, will consist of a core site with a transmitting and receiving radar arrays and four sites with receiving antenna arrays at some 100 km from the core. The fully operational 5-site system will generate several thousand times data of current EISCAT system with 40 PB/year in 2022 and is expected to operate for 30 years. EISCAT 3D data e-Infrastructure plans to use the high performance computers for central site data processing and high throughput computers for mirror sites data processing. Downloading the full data is not time critical, but operations require real-time information about certain pre-defined events to be sent from the sites to the operation center and a real-time link from the operation center to the sites to set the mode of radar operation on with immediate action. See figure in full description.

2.42 ENVRI, Common Operations of Environmental Research Infrastructure; Yin Chen, Cardiff University

Application: ENVRI Research Infrastructures (ENV RIs) addresses European distributed, long-term, remote controlled observational networks focused on understanding processes, trends, thresholds, interactions and feedbacks and increasing the predictive power to address future environmental challenges. ENVRI includes ICOS is a European
distributed infrastructure dedicated to the monitoring of greenhouse gases (GHG) through its atmospheric, ecosystem and ocean networks; EURO-Argo is the European contribution to Argo, which is a global ocean observing system; EISCAT-3D separately described is a European new-generation incoherent-scatter research radar for upper atmospheric science; LifeWatch separately described is an e-science Infrastructure for biodiversity and ecosystem research; EPOS is a European Research Infrastructure on earthquakes, volcanoes, surface dynamics and tectonics; EMSO is a European network of seafloor observatories for the long-term monitoring of environmental processes related to ecosystems, climate change and geo-hazards; IAGOS Aircraft for global observing system; SIOS Svalbard arctic Earth observing system

Current Approach: ENVRI develops a Reference Model (ENVRI RM) as a common ontological framework and standard for the description and characterization of computational and storage infrastructures in order to achieve seamless interoperability between the heterogeneous resources of different infrastructures. The ENVRI RM serves as a common language for community communication, providing a uniform framework into which the infrastructure’s components can be classified and compared, also serving to identify common solutions to common problems. Note data sizes in a given infrastructure vary from gigabytes to petabytes per year.

Futures: ENVRI’s common environment will empower the users of the collaborating environmental research infrastructures and enable multidisciplinary scientists to access, study and correlate data from multiple domains for "system level" research. It provides Bigdata requirements coming from interdisciplinary research.

As shown in Figure 1 in full description, analysis of the computational characteristics of the 6 ESFRI Environmental Research infrastructure, 5 common subsystems has been identified. The definition of them are given in the ENVRI Reference Model, www.envri.eu/rm:

- **Data acquisition**: collects raw data from sensor arrays, various instruments, or human observers, and brings the measurements (data streams) into the system.

- **Data curation**: facilitates quality control and preservation of scientific data. It is typically operated at a data centre.

- **Data access**: enables discovery and retrieval of data housed in data resources managed by a data curation subsystem.

- **Data processing**: aggregates the data from various resources and provides computational capabilities and capacities for conducting data analysis and scientific experiments.

- **Community support**: manages, controls and tracks users’ activities and supports users to conduct their roles in communities.

As shown in Figure 2 of full description, the 5 sub-system map well to the architectures of the ESFRI Environmental Research Infrastructures.
Figure 1: ENVRI Common Subsystems

Figure 2: Architectures of the ESFRI Environmental Research Infrastructures (A) ICOS Architecture

Figure 2: Architectures of the ESFRI Environmental Research Infrastructures (B) LifeWatch Architecture
Figure 2: Architectures of the ESFRI Environmental Research Infrastructures (C) EMSO Architecture

Figure 2: Architectures of the ESFRI Environmental Research Infrastructures (D) Eura-Argo Architecture

Figure 2: Architectures of the ESFRI Environmental Research Infrastructures (E) EISCAT 3D Architecture
2.43 Radar Data Analysis for CReSIS Remote Sensing of Ice Sheets; Geoffrey Fox, Indiana University

**Application:** This data feeds into intergovernmental Panel on Climate Change (IPCC) and uses custom radars to measures ice sheet bed depths and (annual) snow layers at the North and South poles and mountainous regions.

**Current Approach:** The initial analysis is currently Matlab signal processing that produces a set of radar images. These cannot be transported from field over Internet and are typically copied to removable few TB disks in the field and flown “home” for detailed analysis. Image understanding tools with some human oversight find the image features (layers) that are stored in a database front-ended by a Geographical Information System. The ice sheet bed depths are used in simulations of glacier flow. The data is taken in “field trips” that each currently gather 50-100 TB of data over a few week period.

**Futures:** An order of magnitude more data (petabyte per mission) is projected with improved instrumentation. Demands of processing increasing field data in an environment with more data but still constrained power budget, suggests low power/performance architectures such as GPU systems.

The full descriptions gives workflows for different parts of problem and pictures of operation.

**Figure 1: Typical Radar Data after analysis**
2.44 UAVSAR Data Processing, Data Product Delivery, and Data Services; Andrea Donnellan and Jay Parker, NASA JPL

**Application:** Synthetic Aperture Radar (SAR) can identify landscape changes caused by seismic activity, landslides, deforestation, vegetation changes and flooding. This is for example used to support earthquake science as well as disaster management. This use case supports the storage, application of image processing and visualization of this geolocated data with angular specification.

**Current Approach:** Data from planes and satellites is processed on NASA computers before being stored after substantial data communication. The data is made public as soon as processed and requires significant curation due to instrumental glitches. The current data size is ~150TB.

**Futures:** The data size would increase dramatically if Earth Radar Mission launched. Clouds are suitable hosts but are not used today in production.
2.45 NASA LARC/GSFC iRODS Federation Testbed; Brandi Quam, NASA Langley Research Center

**Application:** NASA Center for Climate Simulation (NCCS) and NASA Atmospheric Science Data Center (ASDC) have complementary data sets, each containing vast amounts of data that is not easily shared and queried. Climate researchers, weather forecasters, instrument teams, and other scientists need to access data from across multiple datasets in order to compare sensor measurements from various instruments, compare sensor measurements to model outputs, calibrate instruments, look for correlations across multiple parameters, etc.

**Current Approach:** The data includes MERRA (described separately) and NASA Clouds and Earth's Radiant Energy System (CERES) EBAF (Energy Balanced And Filled)-TOA (Top of Atmosphere) Product which is about 420MB and Data from the EBAF-Surface Product which is about 690MB. Data grows with each version update (about every six months). To analyze, visualize and otherwise process data from heterogeneous datasets is currently a time consuming effort that requires scientists to separately access, search for, and download data from multiple servers and often the data is duplicated without an understanding of the authoritative source. Often times accessing data is greater than scientific analysis time. Current datasets are hosted on modest size (144 to 576 cores) Infiniband clusters.

**Futures:** The improved access will be enabled through the use of iRODS that enables parallel downloads of datasets from selected replica servers that can be geographically dispersed, but still accessible by users worldwide. iRODS operation will be enhanced with semantically organized metadata, and managed via a highly precise Earth Science ontology. Cloud solutions will also be explored.

2.46 MERRA Analytic Services MERRA/AS; John L. Schnase & Daniel Q. Duffy, NASA Goddard Space Flight Center

**Application:** This application produces global temporally and spatially consistent syntheses of 26 key climate variables by combining numerical simulations with observational data. Three-dimensional results are produced every 6-hours extending from 1979-present. This supports important applications like Intergovernmental Panel on Climate Change (IPCC) research and the NASA/Department of Interior RECOVER wildfire decision support system; these applications typically involve integration of MERRA with other datasets. The full description has a typical MERRA/AS output.

**Current Approach:** MapReduce is used to process a current total of 480TB. The current system is hosted on a 36 node Infiniband cluster

**Futures:** Clouds are being investigated. The data is growing by one TB a month.
2.47 Atmospheric Turbulence - Event Discovery and Predictive Analytics; Michael Seablom, NASA HQ

**Application:** This builds datamining on top of reanalysis products including the North American Regional Reanalysis (NARR) and the Modern-Era Retrospective-Analysis for Research (MERRA) from NASA where latter described earlier. The analytics correlate aircraft reports of turbulence (either from pilot reports or from automated aircraft measurements of eddy dissipation rates) with recently completed atmospheric re-analyses. This is of value to aviation industry and to weather forecasters. There are no standards for re-analysis products complicating system where MapReduce is being investigated. The reanalysis data is hundreds of terabytes and slowly updated whereas turbulence is smaller in size and implemented as a streaming service. A typical turbulent wave image is in full description.

**Current Approach:** Current 200TB dataset can be analyzed with MapReduce or the like using SciDB or other scientific database.

**Futures:** The dataset will reach 500TB in 5 years. The initial turbulence case can be extended to other ocean/atmosphere phenomena but the analytics would be different in each case.
2.48 Climate Studies using the Community Earth System Model at DOE’s NERSC center; Warren Washington, NCAR

Application: We need to understand and quantify contributions of natural and anthropogenic-induced patterns of climate variability and change in the 20th and 21st centuries by means of simulations with the Community Earth System Model (CESM). The results of supercomputer simulations across the world need to be stored and compared.

Current Approach: The Earth Systems Grid (ESG) enables world wide access to Peta/Exa-scale climate science data with multiple petabytes of data at dozens of federated sites worldwide. The ESG is recognized as the leading infrastructure for the management and access of large distributed data volumes for climate change research. It supports the Coupled Model Intercomparison Project (CMIP), whose protocols enable the periodic assessments carried out by the Intergovernmental Panel on Climate Change (IPCC).

Futures: Rapid growth of data with 30 PB produced at NERSC (assuming 15 end-to-end climate change experiments) in 2017 and many times more this worldwide.

2.49 DOE-BER Subsurface Biogeochemistry Scientific Focus Area; Deb Agarwal, LBNL

Application: Development of a Genome-Enabled Watershed Simulation Capability (GEWaSC) that will provide a predictive framework for understanding how genomic information stored in a subsurface microbiome, affects biogeochemical watershed functioning; how watershed-scale processes affect microbial functioning; and how these interactions co-evolve.

Current Approach: Current modeling capabilities can represent processes occurring over an impressive range of scales (ranging from a single bacterial cell to that of a contaminant plume). Data crosses all scales from genomics of the microbes in the soil to watershed hydro-biogeochemistry. Data are generated by the different research areas and
include simulation data, field data (hydrological, geochemical, geophysical), ‘omics data, and observations from laboratory experiments.

**Futures:** Little effort to date has been devoted to developing a framework for systematically connecting scales, as is needed to identify key controls and to simulate important feedbacks. GEWaSC will develop a simulation framework that formally scales from genomes to watersheds and will synthesize diverse and disparate field, laboratory, and simulation datasets across different semantic, spatial, and temporal scales.

### 2.50 DOE-BER AmeriFlux and FLUXNET Networks; Deb Agarwal, LBNL

**Application:** AmeriFlux and FLUXNET are US and world collections respectively of sensors that observe trace gas fluxes (CO2, water vapor) across a broad spectrum of times (hours, days, seasons, years, and decades) and space. Moreover, such datasets provide the crucial linkages among organisms, ecosystems, and process-scale studies—at climate-relevant scales of landscapes, regions, and continents—for incorporation into biogeochemical and climate models.

**Current Approach:** Software includes EddyPro, Custom analysis software, R, python, neural networks, Matlab. There are ~150 towers in AmeriFlux and over 500 towers distributed globally collecting flux measurements.

**Futures:** Field experiment data taking would be improved by access to existing data and automated entry of new data via mobile devices. Need to support interdisciplinary studies integrating diverse data sources.

### Energy

### 2.51 Consumption forecasting in Smart Grids; Yogesh Simmhan, University of Southern California

**Application:** Predict energy consumption for customers, transformers, sub-stations and the electrical grid service area using smart meters providing measurements every 15-mins at the granularity of individual consumers within the service area of smart power utilities. Combine Head-end of smart meters (distributed), Utility databases (Customer Information, Network topology; centralized), US Census data (distributed), NOAA weather data (distributed), Micro-grid building information system (centralized), Micro-grid sensor network (distributed). This generalizes to real-time data-driven analytics for time series from cyber physical systems

**Current Approach:** GIS based visualization. Data is around 4 TB a year for a city with 1.4M sensors in Los Angeles. Uses R/Matlab, Weka, Hadoop software. Significant privacy issues requiring anonymization by aggregation. Combine real time and historic data with machine learning for predicting consumption.

**Futures:** Wide spread deployment of Smart Grids with new analytics integrating diverse data and supporting curtailment requests. Mobile applications for client interactions.

**NEXT**