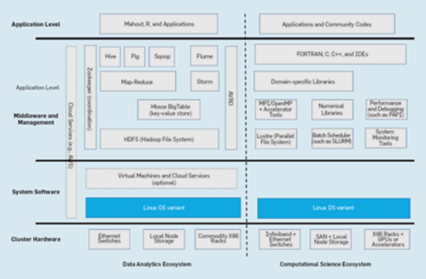
SAND2016-4945PE

Exceptional service in the national interest

Sandia National Laboratories



Application Level | Mahout, R, and Applications | Applications and Community Codes

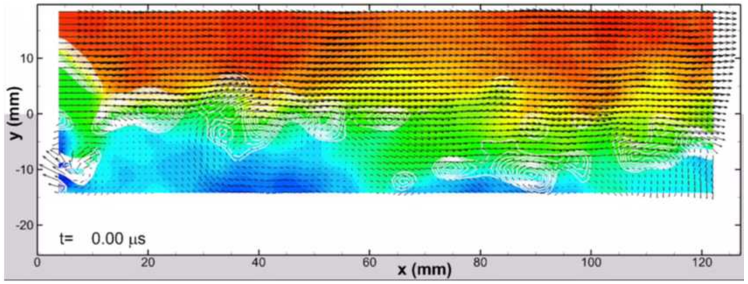
Middleware and Management | Application Level | Cloud Services (e.g., AWS) | Zookeeper (coordination) | Hive | Pig | Sqoop | Flume | Map-Reduce | Storm | Hbase BigTable (key-value store) | HDFS (Hadoop File System) | AVRO | FORTRAN, C, C++, and IDEs | Domain-specific Libraries | MPI/OpenMP + Accelerator Tools | Numerical Libraries | Performance and Debugging (such as PAPI) | Lustre (Parallel File System) | Batch Scheduler (such as SLURM) | System Monitoring Tools

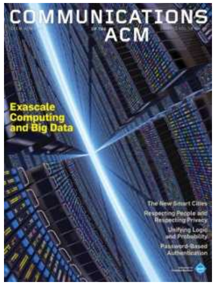
System Software | Virtual Machines and Cloud Services (optional) | Linux OS variant | Linux OS variant

Cluster Hardware | Ethernet Switches | Local Node Storage | Commodity X86 Racks | Infiniband + Ethernet Switches | SAN + Local Node Storage | X86 Racks + GPUs or Accelerators

Data Analytics Ecosystem

Computational Science Ecosystem





COMMUNICATIONS OF THE ACM

Exascale Computing and Big Data

The New Smart Cities

Respecting People and Respecting Privacy

Unifying Logic and Probability

Password-Based Authentication

LAUNCHING IN 2015

IEEE TRANSACTIONS ON BIG DATA

The IEEE Transactions on Big Data (TBD), publishes peer reviewed articles with big data as the main focus. The articles will provide cross disciplinary innovative research ideas and applications results for big data including novel theory, algorithms and applications. Research areas for big data include, but are not restricted to, big data analytics, big data visualization, big data curation and management, big data semantics, big data infrastructure, big data standards, big data performance analyses, intelligence from big data, scientific discovery from big data security, privacy, and legal issues specific to big data. Applications of big data in the fields of endeavor where massive data is generated are of particular interest.

TBD EDITOR-IN-CHIEF:
Qiang Yang, Huawei Noah's Ark Research Lab and the Hong Kong University of Science and Technology

TBD Steering Committee Members

IEEE COMPUTER SOCIETY
Stephen Gago (Chair)
Divyakant (Divy) Agrawal
Vassilis Tsotras

IEEE COMMUNICATIONS SOCIETY
Shuguang (Robert) Cui
[illegible] (Tony) Xia

IEEE COMPUTATIONAL INTELLIGENCE SOCIETY
Nikhil R. Pal

IEEE SENSORS COUNCIL
Mahmoud Daneshmand

IEEE SIGNAL PROCESSING SOCIETY
Malcolm Slaney

IEEE SYSTEMS, MAN, AND CYBERNETICS SOCIETY
Philip Chen

IEEE CONSUMER ELECTRONICS SOCIETY
Stefan Mozar

IEEE SYSTEMS COUNCIL
Paolo Carbone

IEEE VEHICULAR TECHNOLOGY SOCIETY
Fabrice Labeau

For more information please visit: http://www.computer.org/tbd

Data Sciences Overview

Engineering Sciences External Review Board

April 13th, 2016

Kim Mish; V&V, UQ, and Credibility Processes

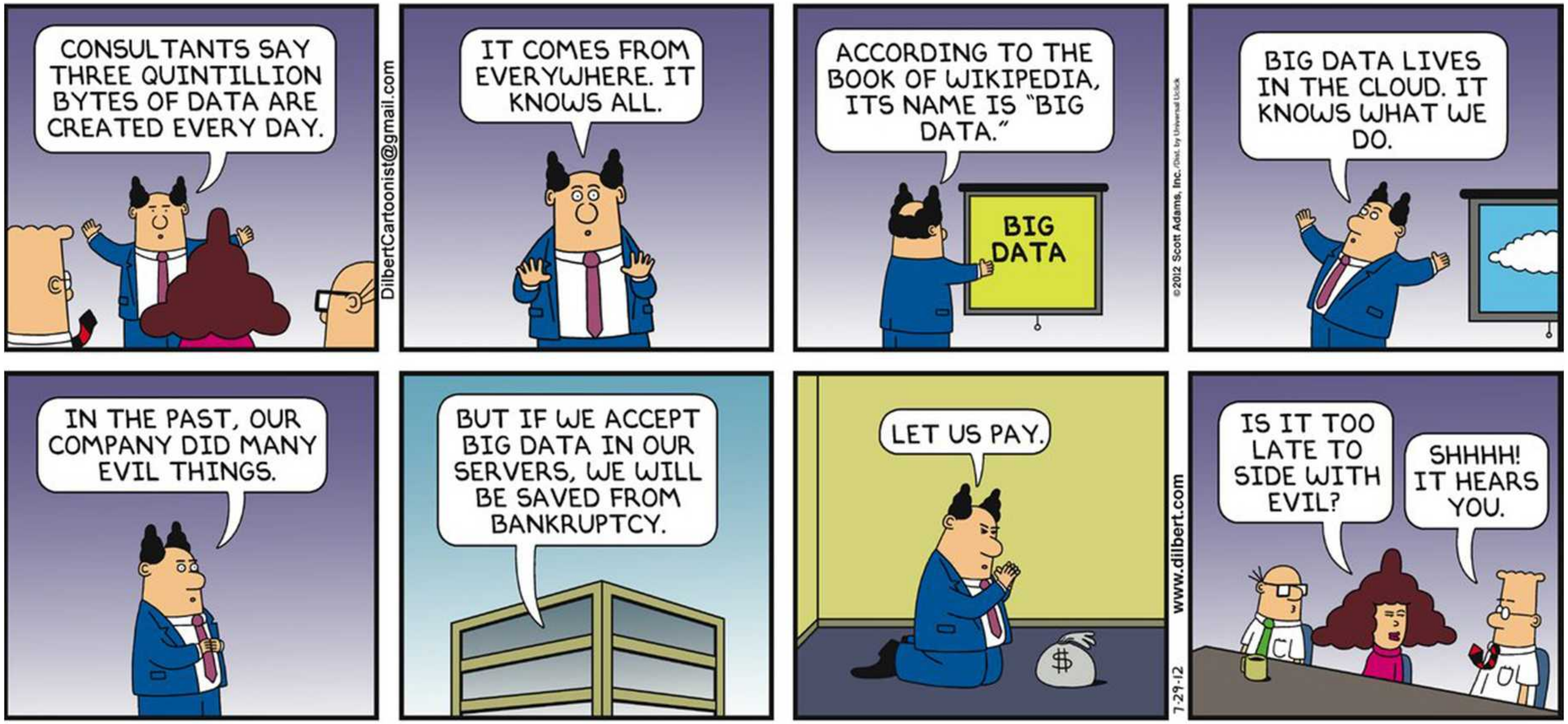
U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. SAND NO. 2011-XXXXP

The Dilbert Technology Test

- **You know that a technology has arrived when it has become the subject of a Dilbert cartoon !!**


CONSULTANTS SAY THREE QUINTILLION BYTES OF DATA ARE CREATED EVERY DAY.
IT COMES FROM EVERYWHERE. IT KNOWS ALL.
ACCORDING TO THE BOOK OF WIKIPEDIA, ITS NAME IS "BIG DATA."
BIG DATA
BIG DATA LIVES IN THE CLOUD. IT KNOWS WHAT WE DO.
IN THE PAST, OUR COMPANY DID MANY EVIL THINGS.
BUT IF WE ACCEPT BIG DATA IN OUR SERVERS, WE WILL BE SAVED FROM BANKRUPTCY.
LET US PAY.
$
IS IT TOO LATE TO SIDE WITH EVIL?
SHHHH! IT HEARS YOU.

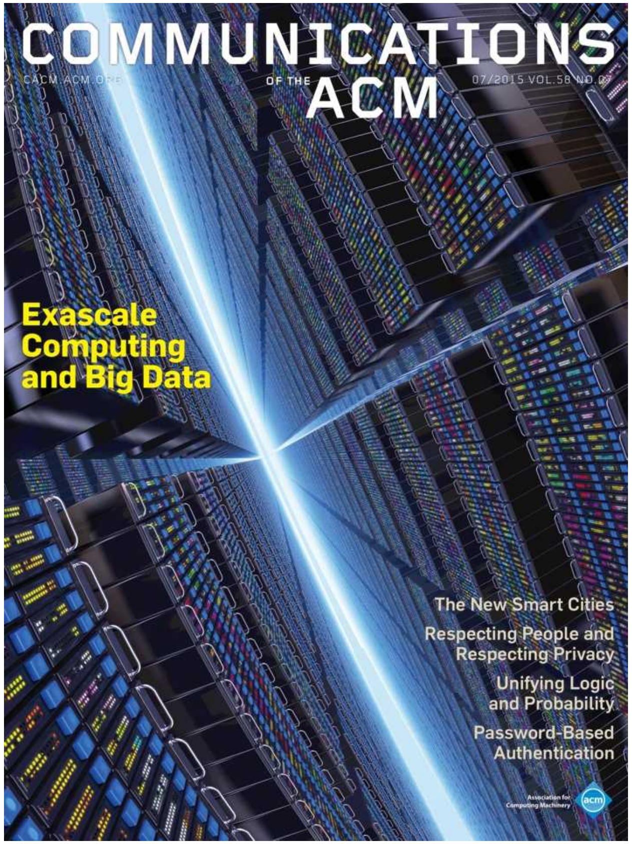
DilbertCartoonist@gmail.com

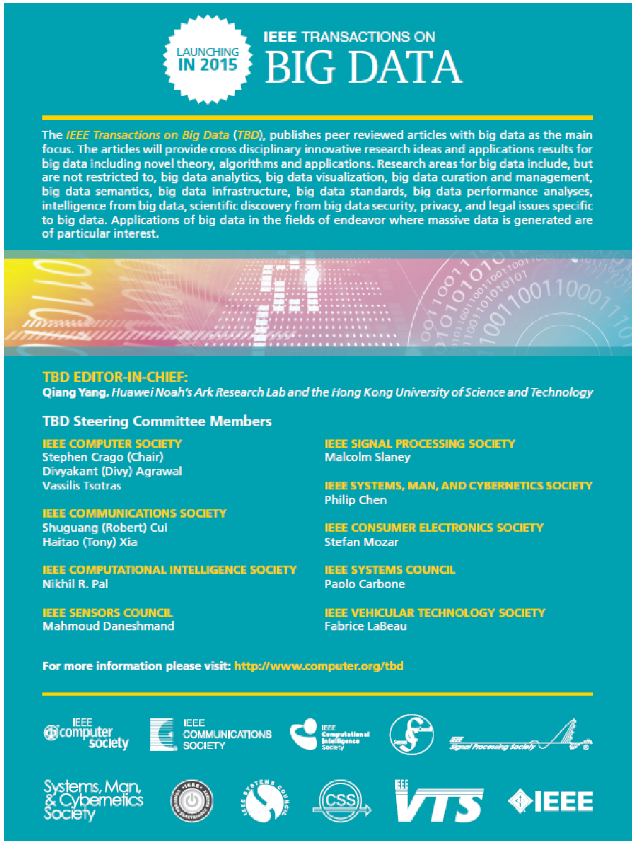
©2012 Scott Adams, Inc./Dist. by Universal Uclick

www.dilbert.com 7-29-12

But Big Data R&D is Here to Stay

- **And both the ACM and the IEEE are producing special volumes and new journals on the subject**


COMMUNICATIONS
OF THE ACM
CACM.ACM.ORG
07/2015 VOL.58 NO.07
Exascale Computing and Big Data
The New Smart Cities
Respecting People and Respecting Privacy
Unifying Logic and Probability
Password-Based Authentication
Association for Computing Machinery
acm


LAUNCHING IN 2015

IEEE TRANSACTIONS ON
BIG DATA

The *IEEE Transactions on Big Data* (*TBD*), publishes peer reviewed articles with big data as the main focus. The articles will provide cross disciplinary innovative research ideas and applications results for big data including novel theory, algorithms and applications. Research areas for big data include, but are not restricted to, big data analytics, big data visualization, big data curation and management, big data semantics, big data infrastructure, big data standards, big data performance analyses, intelligence from big data, scientific discovery from big data security, privacy, and legal issues specific to big data. Applications of big data in the fields of endeavor where massive data is generated are of particular interest.

TBD EDITOR-IN-CHIEF:
Qiang Yang, *Huawei Noah's Ark Research Lab and the Hong Kong University of Science and Technology*

TBD Steering Committee Members

IEEE COMPUTER SOCIETY
Stephen Crago (Chair)
Divyakant (Divy) Agrawal
Vassilis Tsotras

IEEE COMMUNICATIONS SOCIETY
Shuguang (Robert) Cui
Haitao (Tony) Xia

IEEE COMPUTATIONAL INTELLIGENCE SOCIETY
Nikhil R. Pal

IEEE SENSORS COUNCIL
Mahmoud Daneshmand

IEEE SIGNAL PROCESSING SOCIETY
Malcolm Slaney

IEEE SYSTEMS, MAN, AND CYBERNETICS SOCIETY
Philip Chen

IEEE CONSUMER ELECTRONICS SOCIETY
Stefan Mozar

IEEE SYSTEMS COUNCIL
Paolo Carbone

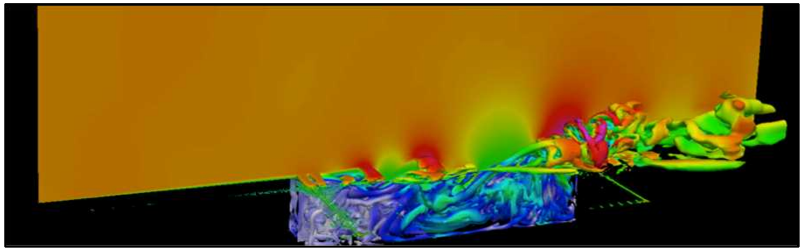
IEEE VEHICULAR TECHNOLOGY SOCIETY
Fabrice LaBeau

For more information please visit: http://www.computer.org/tbd

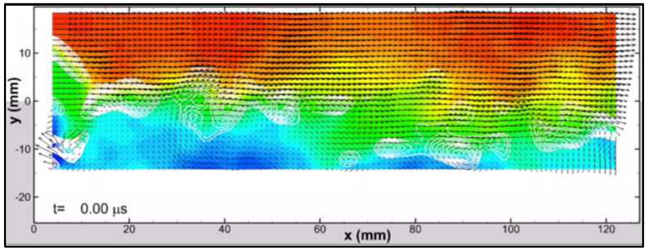
IEEE computer society
IEEE COMMUNICATIONS SOCIETY
IEEE Computational Intelligence Society
Systems, Man, & Cybernetics Society
CSS
VTS
IEEE

Our Mission Meets Data Science

- **Generation, movement, fusion, and archival of data has become of paramount importance at Sandia**
  - **Laboratory has initiated three data sciences campaigns**
    - **Overseen by the new Data Sciences Leadership Team (DSLT)**
    - **Geospatial Imaging Research Challenge (mission focus)**
    - **Streaming Data Research Campaign (mission focus)**
    - **Simulation and Experimental Data Analysis Campaign (ES focus)**
      - Includes NW mission sensibilities, along with climate and energy foci
      - ES partnering with Computational and Information Sciences (CIS)
      - Includes the fusion of field, experimental, and computational data
  - **Our fundamental need in two pictures:**
    - **We are up to our ears in big data, from both real-world and virtual realms**



This is a computational simulation



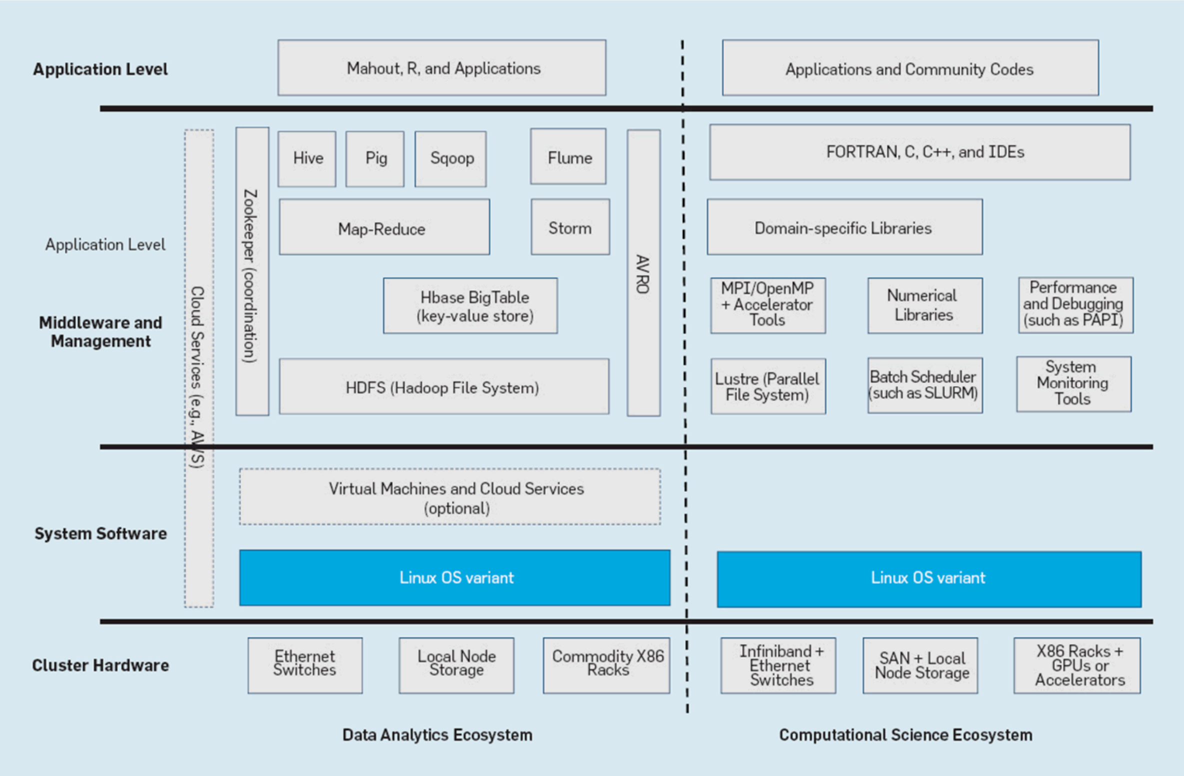
This is an experiment

Sandia Data Strength and Weakness

- The good news
  - Pockets of world-class R&D exist at the laboratory
    - Broad expertise in virtually all areas of big-data R&D
      - Primarily found in mission organizations (e.g., Defense and Systems Assessments Division), with some support from ES and other centers
    - Lab has thus formed strong intellectual cadre of SMEs
      - Many of these are members of the lab's DSLT
- The not-so-good news
  - Expertise is exactly that: it is found in pockets!
    - Need to develop a broader lab-wide strategy for data R&D
    - ES is appropriate venue for this lab-wide education
      - ESRF-related organizations help form the S&T basis for Sandia
      - I am the chair of the ES education council, so well-poised for success here
- But here's some more good news
  - Scalable data sciences is closely related to scalable HPC
    - And ES/CIS organizations are very, very good at scalable HPC

Reed & Dongarra on Big-Data HPC

- **A picture is worth a thousand words: by leveraging our HPC expertise, we can show leadership in big-data R&D**



Application Level

Mahout, R, and Applications

Applications and Community Codes

Application Level

Middleware and Management

Cloud Services (e.g., AWS)

Zookeeper (coordination)

Hive

Pig

Sqoop

Flume

Map-Reduce

Storm

Hbase BigTable (key-value store)

HDFS (Hadoop File System)

AVRO

FORTRAN, C, C++, and IDEs

Domain-specific Libraries

MPI/OpenMP + Accelerator Tools

Numerical Libraries

Performance and Debugging (such as PAPI)

Lustre (Parallel File System)

Batch Scheduler (such as SLURM)

System Monitoring Tools

System Software

Virtual Machines and Cloud Services (optional)

Linux OS variant

Linux OS variant

Cluster Hardware

Ethernet Switches

Local Node Storage

Commodity X86 Racks

Infiniband + Ethernet Switches

SAN + Local Node Storage

X86 Racks + GPUs or Accelerators

Data Analytics Ecosystem

Computational Science Ecosystem

Don't need to examine details here: just note that the left column (data analytics) has much in common with the right column (computational science)

Daniel A. Reed and Jack Dongarra
Communications of the ACM, Vol. 58 No. 7, Pages 56-68

Towards a Data Sciences Community

- Our objective is to build a community of data-science practitioners from within our ranks
- Implementation plan has three motivating goals:
  - Goal #1: Consciousness-Raising
    - Bring in experts from outside the lab to help us learn and grow
    - Become part of the national community in data sciences
  - Goal #2: Grow Our Data Sciences Teams
    - Build on success of internal projects (including LDRDs)
    - Fill in gaps as needed via help from Goal #1, e.g., repositories
  - Goal #3: Outreach to Other Agencies
    - We are part of a federal community of practice in data sciences
    - Seeking outreach opportunities with academic and agency partners

Goal #1: Consciousness-Raising

- Lab-wide educational efforts for FY2016
  - These are proposed as near-term activities (2 or 3 years)
  - Track 1: Seminar series
    - External speakers brought here to raise consciousness about data sciences in general, and big data in particular
      - Purpose is to bring in speakers who have strengths in the fundamentals of data science and in their practical application
  - Track 2: Community-building workshops
    - Short (1-day) workshops on topics with a Sandia-specific focus
      - Purpose is to consolidate the growth of expertise in data science within the laboratory community
      - Topics include DIC, repository design, and data fusion
  - And there's more, too, including speed-dating!
    - LDRD PI event: interact with lab's data science PIs
    - Official Title/Date: LDRD PI Roundtable, April 25th, 2016

Initial FY2016 Seminar Speakers

Vipin Kumar (May)

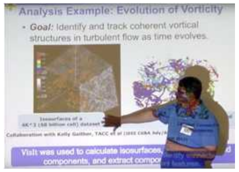
Topic: Fundamentals of Scalable Data Science

- William Norris Chair in Computer Science & Engineering, University of Minnesota
- Author/co-author of over 300 research articles and over 10 books, including “Introduction to Parallel Computing” and “Introduction to Data Mining”
- Director (1998-2005) of DoD HPC Research Center

Cherri Pancake (June)

Topic: Usability of Large Data Archives

- Professor of Electrical Engineering & Computer Science, Oregon State University
- Intel Faculty Fellow and Director of NACSE (Northwest Alliance for Computational Science and Engineering)
- Fellow of both the ACM and the IEEE


Analysis Example: Evolution of Vorticity
• Goal: Identify and track coherent vortical structures in turbulent flow as time evolves.
Isosurfaces of a 4K^3 (68 billion cell) dataset
Collaboration with Kelly Gaither, TACC et al
VisIt was used to calculate isosurfaces,
components, and extract compo

Hank Childs (July)

Topic: Exascale Visualization

- Associate Professor in Computer and Information Science Department, University of Oregon
- Architect of Lawrence Livermore’s VISIT project, a highly-scalable visual analysis tool used worldwide
- Over a decade of experience managing big-data visualization projects within the DOE complex

Hank Jenkins-Smith and Carol Silva (Aug)

Topic: Tools for Big-Data Policy Analysis

- Professors and Co-Directors, Center for Energy, Security, and Society at the University of Oklahoma
- Long history of policy collaboration with Sandia’s nuclear power R&D organizations
- Specialization in textual analysis of document streams affecting public policy decisions

Goal #2: Grow Data Sciences Teams

- Start with existing projects that are leading-edge R&D activities, using these as exemplars
  - Essential to find these successful examples, and to help them become the nucleus of a long-term R&D initiative
  - Will see two examples today:
    - Matt Barone: Turbulent Flow UQ Using ML Techniques
    - Ed Jimenez: Multi-Energy Iterative Volumetric Reconstruction
      - Presented by Elizabeth Lopez due to travel constraints
  - Many others exist that will not be presented, including:
    - Philip Reu: Digital Image Correlation LDRD
    - “Born Qualified” Grand Challenge LDRD
    - Topological optimization R&D activities
    - Hardening of inverse methods technologies
  - In contrast with Goal #1, this is a long-term enterprise

Goal #3: Outreach to Other Agencies

- **President's federal big-data initiative provides opportunities for interagency cooperation**
  - **DoD CREATE All-Hands (May 3rd-5th)**
    - **CREATE program is the DoD's ASC equivalent**
      - **We have been invited to brief DoD on SNL big-data initiative and to host a session on big-data applications**
      - **Ultimate goal is a DoD/DOE collaboration on big data**
  - **Outreach to federal agencies**
    - **Visits in works for FY16/17 to NSF, NASA, NOAA and others**
    - **Include both exploratory and mission agencies in outreach**
    - **DOE is lead agency for HPC under NSCI (National Strategic Computing Initiative)**
  - **Outreach to universities**
    - **Laboratory's new Academic Alliance program is one example**

Summary

- **We are in the data sciences business**
  - **Data sciences is a cross-cutting technology that underlies virtually all fields of science and engineering**
    - **Dilbert, ACM, and IEEE agree: big data is here to stay**
  - **Sandia possesses a unique combination of physical resources and leading-edge computational resources**
    - **Sandia is indeed a "national laboratory", with an incredible variety of world-class laboratory facilities, each serving our mission as an instrument of national policy**
    - **Sandia and its cousin laboratories in the DOE complex lead the federal effort in next-generation supercomputing**
  - **All that is missing at present is for Sandia to join the larger community of scholarship on data sciences**
    - **We hope that the ESRF review panel can help us with this!**