



U.S. DEPARTMENT OF
ENERGY

Office of
Science

High Priority Investments for Data-Intensive Science

June 17, 2014

Big Data Symposium

Arlington, VA

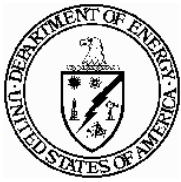
Ceren Susut, PhD
Program Manager

Advanced Scientific Computing Research (ASCR)

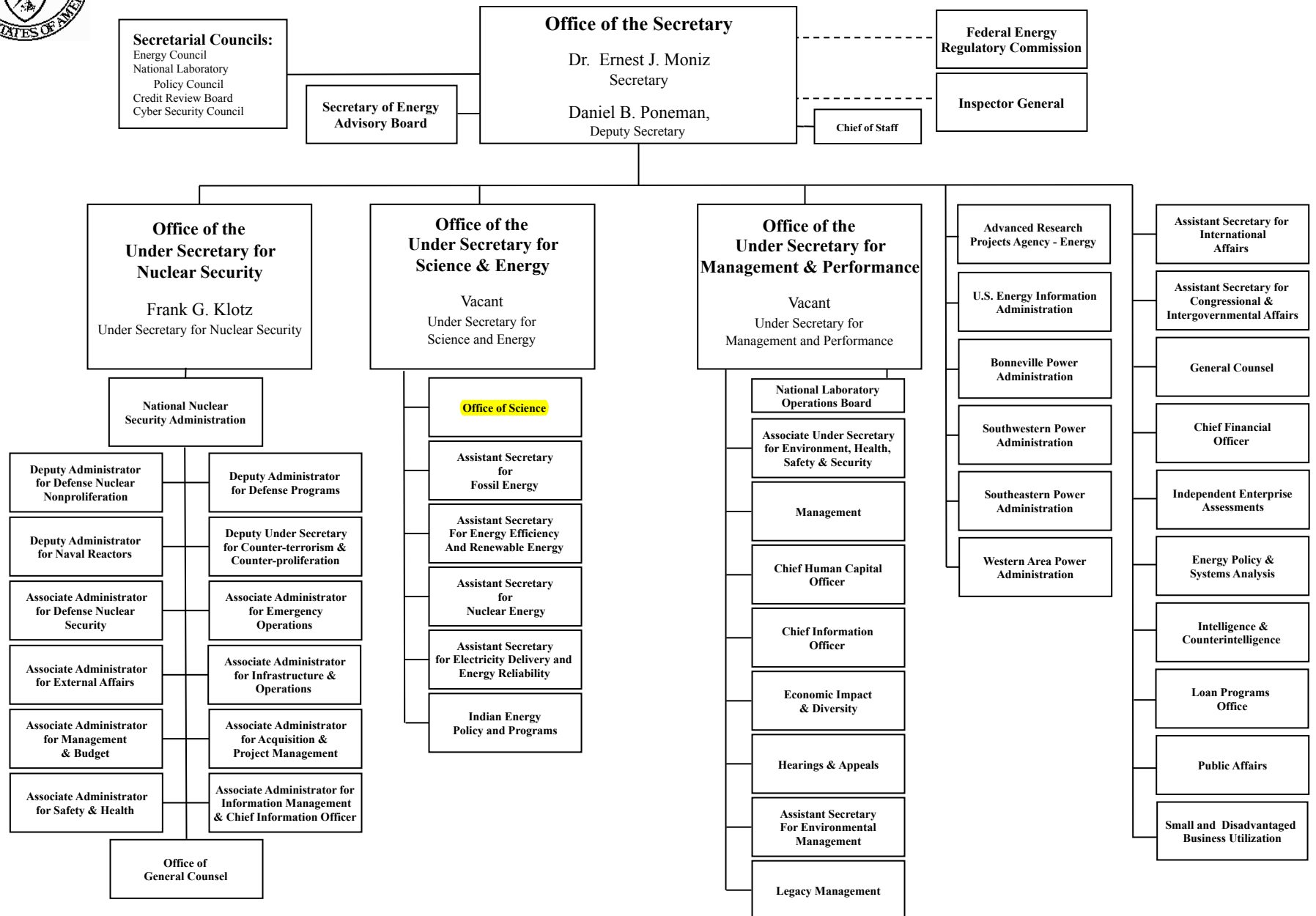
Office of Science

U.S. Department of Energy

<http://science.energy.gov/ascr>



DEPARTMENT OF ENERGY

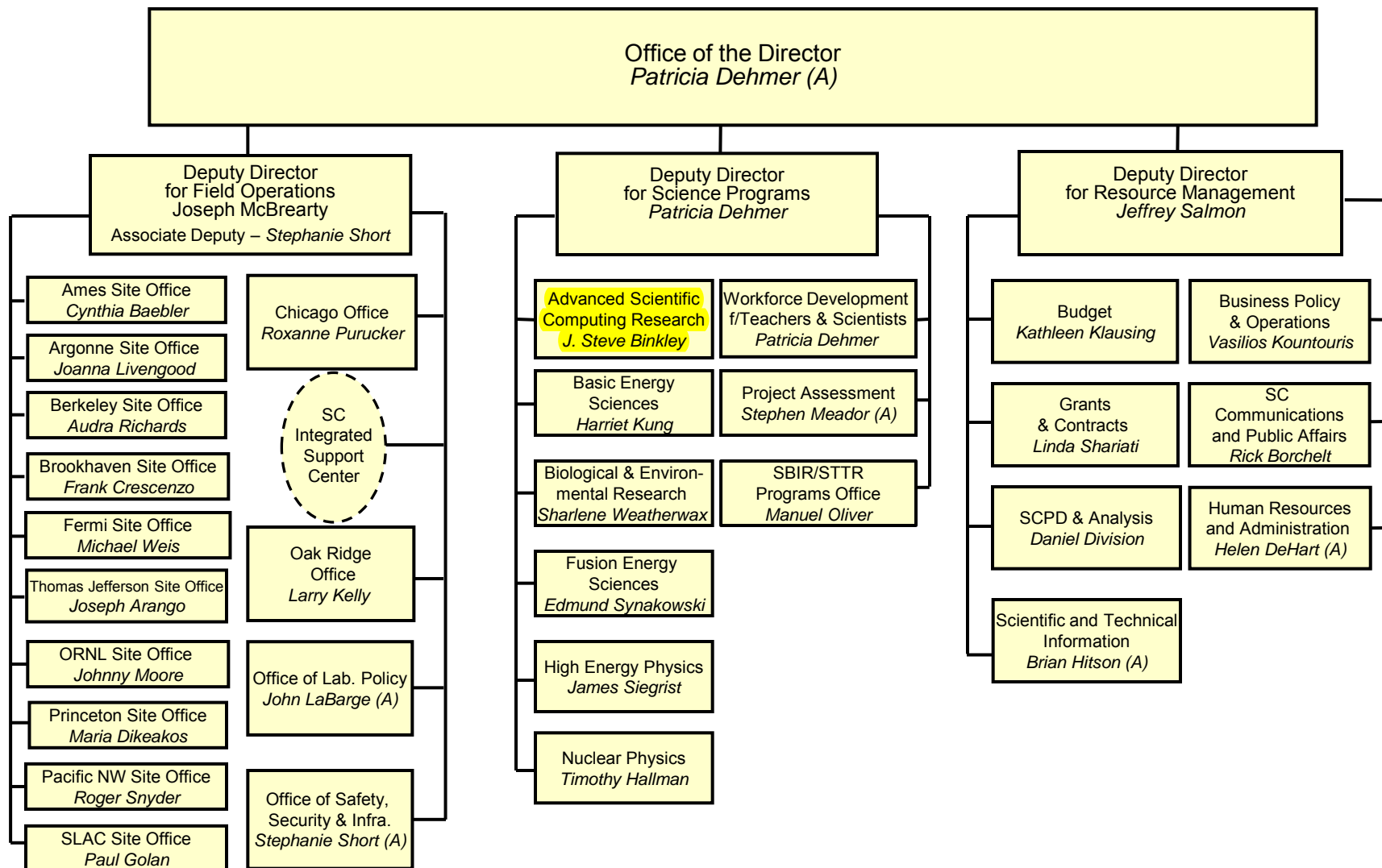


DOE's Office of Science, by the Numbers



- \$5B budget, supporting:
 - 25,000 Ph.D. scientists, graduate students, undergraduates, engineers, and technical staff at more than 300 institutions in all 50 States and DC through competitive awards
 - 26,000 users at 32 national scientific user facilities
- 100 Nobel Prizes during the past 6 decades—more than 20 in the past 10 years

The undulator hall at the
Linac Coherent Light Source, SLAC, 2011.



Advanced Scientific Computing Research (ASCR)

Mission

Discover, develop, and deploy computational and networking capabilities to analyze, model, simulate, and predict complex phenomena important to the Department of Energy (DOE).

- Partnerships to Transform the Present: Mutually beneficial collaborations to dramatically accelerate progress in scientific computing
- Research to Enable the Future: Advance applied mathematics, computer science and high-performance networks
- World-class Facilities: Supercomputing facilities and advanced scientific networks; Argonne and Oak Ridge Leadership Computing Facilities, NERSC, ESNet

(dollars in thousands)

FY 2013 Current	FY 2014 Current	FY2015 President's Request
405,000	478,093	541,000



The DOE/SC Labs Today – User Facilities



Berkeley, California
202 acres and 106 buildings
3,400 FTEs
1,084 students & postdocs
8,579 facility users

www.lbl.gov



Pacific Northwest
NATIONAL LABORATORY



Richland, Washington
600 acres and 101 buildings
4,180 FTEs
567 students & postdocs
2,414 facility users

www.pnnl.gov

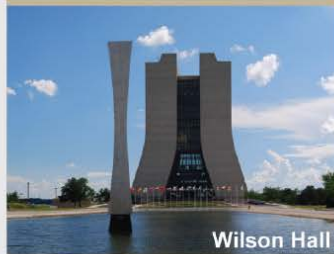


THE Ames Laboratory



Ames, Iowa
10 acres and 12 buildings
315 FTEs
210 students & postdocs

www.ameslab.gov



Wilson Hall

Batavia, Illinois
6,800 acres and 356 buildings
1,914 FTEs
1,090 students & postdocs
2,317 facility users

www.fnal.gov



Advanced Photon Source

Argonne, Illinois
1,500 acres and 99 buildings
3,375 FTEs
1,147 students & postdocs
4,289 facility users

www.anl.gov



Menlo Park, California
426 acres and 142 buildings
1,681 FTEs
300 students & postdocs
3,384 facility users

www.slac.stanford.edu



Spallation Neutron Source

Oak Ridge, Tennessee
4,470 acres and 252 buildings
4,533 FTEs
1,753 students & postdocs
3,116 facility users

www.ornl.gov



Newport News, Virginia
169 acres and 63 buildings
769 FTEs
74 students & postdocs
1,376 facility users

www.jlab.org



NSTX Spherical Tokamak

Princeton, New Jersey
89 acres and 34 buildings
428 FTEs
66 students & postdocs
145 facility users

www.pppl.gov



Relativistic Heavy Ion Collider

Upton, New York
5,320 acres and 331 buildings
2,990 FTEs
593 students & postdocs
4,253 facility users

www.bnl.gov

Data Intensive Science



Genomics

Data Volume increases
to 10 PB in FY21



High Energy Physics (Large Hadron Collider)

15 PB of data/year



Light Sources

Approximately
300 TB/day



Climate

Data expected to be
hundreds of 100 EB

Driven by exponential technology advances

Data sources

- Advanced Experimental and Observational Facilities
- Scientific Computing Facilities
- Combination of simulation/observational/experimental data

Challenges

- Volume/velocity
- Heterogeneity
- Complexity
- Integration
- Timeliness
- Cost



Data drivers from DOE Scientific User Facilities†

	2013 Current data rate*	2015 Projected need	2018 Projected need
HEP Cosmic Frontier example – Large Synoptic Survey Telescope	~0.2 GB/s	~0.5 GB/s	~1-10 GB/s
HEP Energy Frontier Example – Atlas LHC	1 GB/s*	2 GB/s*	4 GB/s*
HEP Intensity Frontier Example – Belle II	1 GB/s	2 GB/s	20 GB/s
BER Climate	100 GB/s	1000 GB/s	1000 GB /s
BER EMSL – one instrument example - TEM	100 – 1000 images (2Kx2K)/ per day	1000 Images/s = 2GB/s	1,000,000 Images/s = 2 TB/s
BER JGI example - Illumina HiSeq	18 MB/s	72 MB/s	600 MB/s
BES Advanced Photon Source example – 2-BM Beamline	1 GB/s/beamline		10 GB/s
BES Nano Science example – X-Ray Spectroscopy		100 MB x 100 excited atoms x 100 snapshots = 1 TB per point (P,T)	
BES Neutron Facilities	~0.05GB/s	~0.10 GB/s	~0.30GB/s

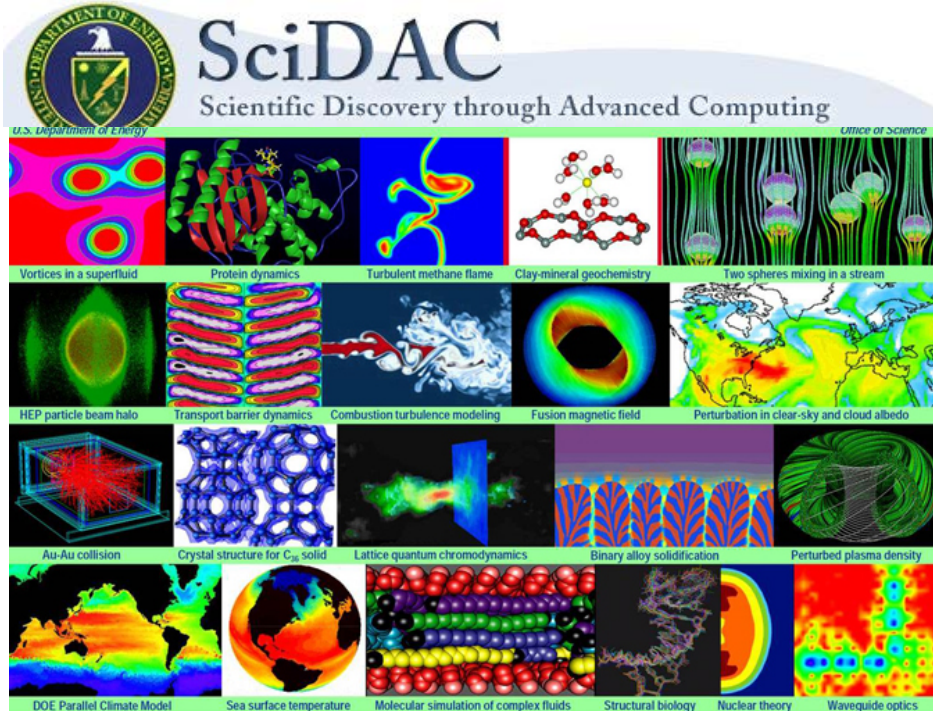
* Data Rate after 99% reduction in hardware data acquisition system

† http://science.energy.gov/~media/ascr/pdf/program-documents/docs/ASKD_Report_V1_0.pdf

Scientific Discovery through Advanced Computing (SciDAC)

Collaborations of applied mathematicians, computer scientists and domain scientists to advance scientific frontiers through modeling, simulation and analysis

Ranked #1 among successful worldwide HPC programs*



- Started in 2001, 5-year projects
- Third generation, 18 partnerships across Office of Science and 4 SciDAC Institutes
- www.scidac.gov



SciDAC Institutes are large team projects involving National Laboratory, University and Industry collaborators

FASTMath Director – Lori Diachin Scalable solvers & discretizations	QUEST Director – Habib Najm Uncertainty Quantification	SDAV Director – Arie Shoshani Scalable data management, analysis & visualization	SUPER Director – Robert Lucas Performance tools & code optimization
Lawrence Livermore (CA)	Sandia (CA)	Lawrence Berkeley (CA)	Univ of Southern CA
Argonne (IL)	Los Alamos (NM)	Argonne (IL)	Argonne (IL)
Lawrence Berkeley (CA)	Duke University (NC)	Lawrence Livermore (CA)	Lawrence Berkeley (CA)
Sandia (CA & NM)	MIT (MA)	Los Alamos (NM)	Lawrence Livermore (CA)
RPI (NY)	Univ of Southern CA	Oak Ridge (TN)	Oak Ridge (TN)
	Univ of Texas, Austin (TX)	Sandia (NM)	Univ of CA, San Diego (CA)
		Univ of CA, Davis (CA)	Univ of Maryland (MD)
		Georgia Tech (GA)	Univ of North Carolina (NC)
		North Carolina St Univ (NC)	Univ of Oregon (OR)
		Northwestern (IL)	Univ of Tenn, Knoxville (TN)
		Ohio State Univ (OH)	Univ of Utah (UT)
		Rutgers Univ (NJ)	
		Univ of Utah (UT)	
		Kitware, Inc (NY)	

FASTMath - Frameworks, Algorithms &
Scalable Technologies for Mathematics

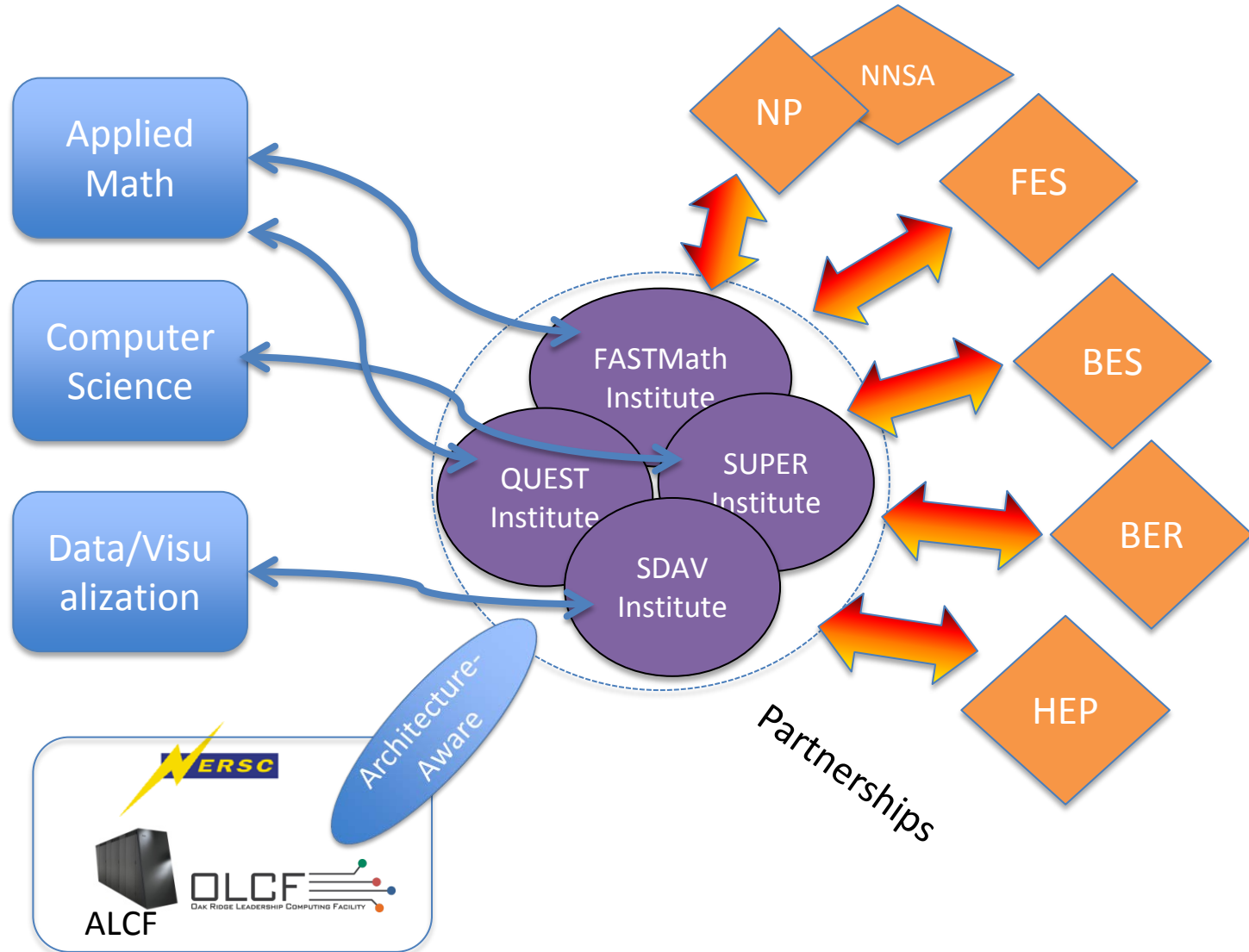
QUEST - Quantification of Uncertainty in
Extreme-Scale Computations

SDAV - Scalable Data Management,
Analysis & Visualization

SUPER - Institute for Sustained
Performance, Energy & Resilience

	Lead
	National Labs
	Universities
	Industry

SciDAC connects ASCR base research with science



SciDAC Data Analysis and Visualization Institute (SDAV)

SDAV Goals:

- to actively work with application teams to assist them in achieving breakthrough science
- to provide technical solutions in the data management, analysis, and visualization regimes that are broadly used by the computational science community running on Leadership Class Machines
- To use existing robust tools to the extent possible and develop/adapt tools on an as-needed basis

Data Management tools

- In Situ Processing and Code Coupling
 - ADIOS
 - Glean
- Indexing
 - FastBit
- In Situ Data Compression
 - ISABELLA
- Parallel I/O and File Formats
 - PnetCDF
 - BP-files
 - HDF5

Data Analysis tools

- Statistical and Data Mining Techniques
 - NU-Minebench
- Importance-Driven Analysis Techniques
 - Domain-Knowledge Directed
 - Geometry Based
- Topological Methods
 - In Situ Topology
 - Feature-Based Analysis
 - High-Frequency Analysis and Tracking

Visualization tools

- Parallel visualization
 - Visit
 - ParaView
- VTK-m framework
- Flow Visualization Methods
- Rendering
- Ensembles, Uncertainty, and Higher-Dimensional Methods



www.SDAV-SciDAC.org



U.S. DEPARTMENT OF
ENERGY

Office of
Science

FastBit and ADIOS

FastBit: Efficient Search Technology for Data Driven Science, quickly finds records satisfying user-specified conditions from a large, complex data set

- Available open source from <http://sdm.lbl.gov/fastbit/> (>10,000 downloads), received 2008 R&D 100 Award
- Gene Context Analysis in IMG used to time-out when comparing 5 or more organisms; with FastBit technology, the hardest version of this problem requires no more than 10 seconds
- Searched through trillion-particle data set from an astronomy application in seconds: “This is the first time anyone has ever queried and visualized 3D particle datasets of this size.” -- Homa Karimabadi, Physicist from UCSD
- Testimonial “FastBit is at least 10x, in many situations 100x, faster than current commercial database technologies” -- Senior Software Engineer, Yahoo! Inc

ADIOS: Adaptable Input/Output System

- Available open source from <https://www.olcf.ornl.gov/center-projects/adios/>, received 2013 R&D 100 Award
- Provides a simple, flexible way for scientists to describe the data in their code that may need to be written, read, or processed outside of the running simulation.



15% More Accuracy in Seasonal Hurricane Forecasts through Comparative Climate Networks Analytics

Nagiza Samatova NCSU/ORNL and Fred Semazzi NCSU

Objectives

- Develop predictive forecasting methodology for climate extremes (e.g., hurricanes, droughts, rainfalls)
- Devise scalable algorithms for predictive mining of large-scale climate complex networks
- Provide mechanistic insights about the key factors contributing to extreme events variability
- Demonstrate high predictive skill for North Atlantic seasonal hurricane activity

Impact

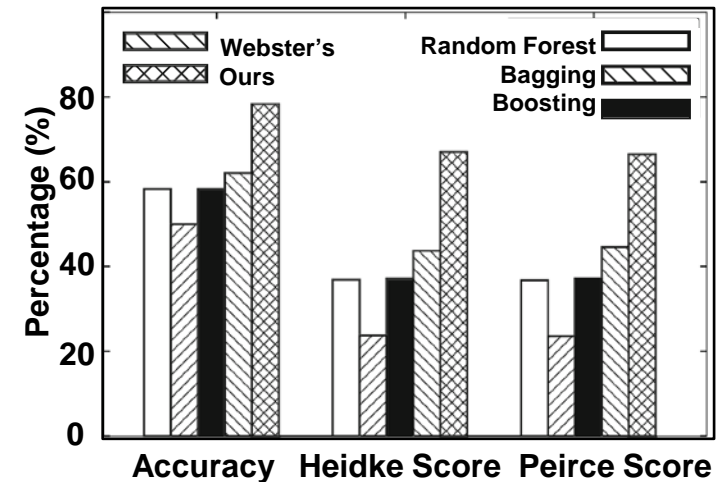
- Provide policy makers more reliable information on seasonal climate extremes
- Scalable large-scale graph mining algorithms of broader applicability (e.g., bioenergy)
- Advance our understanding of the mechanisms that influence hurricane variability and behavior
- International impact managing meningitis epidemic outbreaks driven by climate extremes

Accomplishments

15 percent more accurate forecast of seasonal hurricane activity
Comparative climate networks analytics & machine learning methods

“Novel data-driven methods promise to excel beyond the traditional methods in climate prediction tools”
(Fred Semazzi, Nobel Prize co-winner, climate scientist)

Z. Chen, W. Hendrix, H. Guan, I. Tetteh, A. Choudhary, F. Semazzi, N. Samatova, “Discovery of extreme events-related communities in contrasting groups of physical system networks,” *Data Mining and Knowledge Discovery*, 27(2), p. 225-258, 2012.



U.S. DEPARTMENT OF
ENERGY

Office of
Science



SDAV

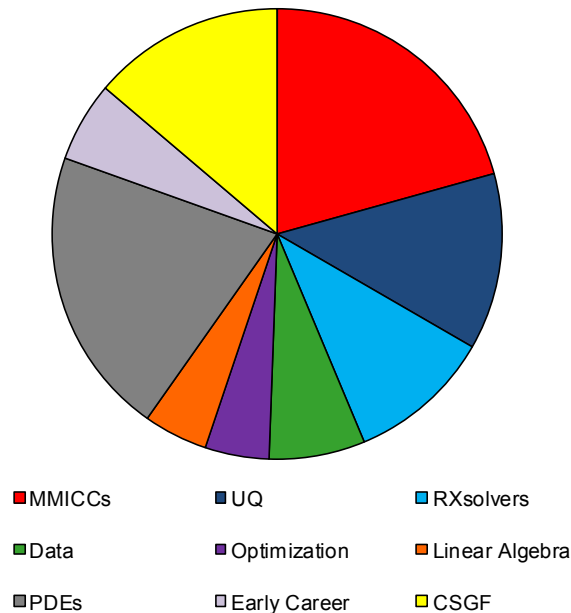
Scalable Data Management, Analysis, and Visualization

Applied Mathematics

Underpins Exascale and Data

Supports the research and development of applied mathematics models, methods and algorithms for understanding natural and engineered systems related to DOE's mission ... with focus on discovery of new applied mathematics, for the ultra-low power, multicore-computing, and data-intensive future.

Research Supported in 2013



- Mathematical Multifaceted Integrated Capability Centers (MMICCs)
- **Uncertainty Quantification**
- **Resilient eXtreme scale Solvers**
- **Mathematics for analysis of extremely large Datasets**
- Optimization
- Advanced linear algebra
- Partial Differential equations (Multiscale mathematics and multiphysics)
- Early Career Research Program
- Computational Science Graduate Fellowship

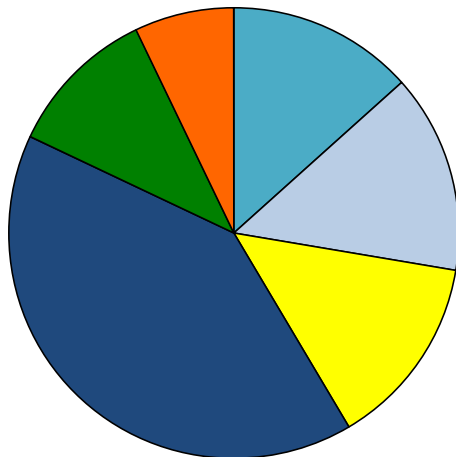


Computer Science

Underpins Exascale and Data

Supports the research and development of today's and tomorrow's leading edge computers and tools for DOE science and engineering and extracting scientific information, discovery and insight from massive data from experiments and simulation...with focus on exascale implications.

Research supported in 2013



- Advanced Architectures
- Extreme Scale Operating and File Systems.
- Performance and Productivity
- Programming Models and Tools
- Data Management and Visualization
- Early Career Research Program



Computer Science 2014 Funding Opportunity Announcement: Scientific Data Management, Analysis & Visualization at Extreme Scale

Basic computer science research that will lay the foundation for building the software infrastructure to support scientific data management, analysis and visualization in the context of extreme scale computing.

Funding Opportunity released in December 2013 totaling \$4M/year for 3 years

- DOE National Laboratory announcement (Lab 14-1043)
http://science.energy.gov/~media/grants/pdf/lab-announcements/2014/LAB_14_1043
- University & industry announcement (DE-FOA-0001043)
http://science.energy.gov/~media/grants/pdf/foas/2014/SC_FOA_0001043.pdf

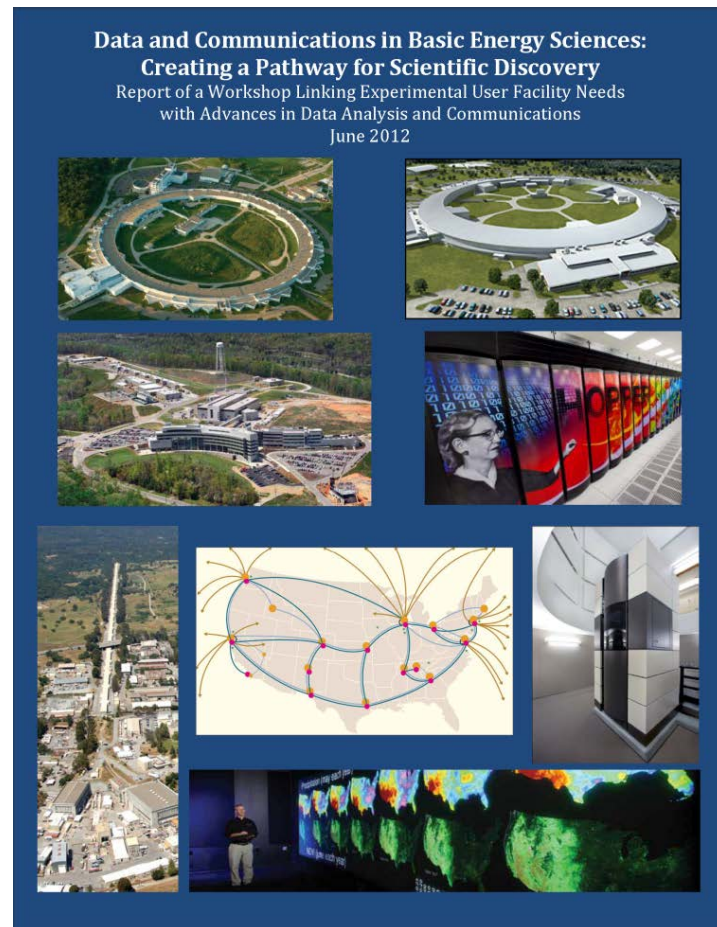
Full proposals were due April 2014. Proposals substantively addressed the following areas:

1. Usability and user interface design;
2. In situ methods for data management, analysis and visualization;
3. Design of in situ workflows to support data management, processing, analysis and visualization;
4. New approaches to scalable interactive visual analytic environments;
5. Proxy applications or workflows and/or simulations for data management, analysis and visualization software to support co-design of extreme scale systems.

Data and Communications in Basic Energy Sciences

- **Workshop Goal:** To identify current and anticipated issues in the acquisition, analysis, communication and storage of experimental data in basic energy sciences and to create the foundation for information exchanges and collaborations among ASCR-BES research and facilities communities to address these issues.
- Focus on high-value-added areas:
 - Integrate theory and analysis components seamlessly within experimental workflow.
 - Move analysis closer to experiment.
 - Match data management access and capabilities with advancements in detectors & sources.
- Continue dialog and collaboration between BES and ASCR on areas of interest and benefit
- Maximize scientific impact from existing and enhanced user facilities

“To jumpstart these activities, there was a strong desire to see a joint effort between ASCR and BES along the lines of the highly successful Scientific Discovery through Advanced Computing program (SciDAC)”



http://science.energy.gov/~media/ascr/pdf/research/scidac/ASCR_BES_Data_Report.pdf



U.S. DEPARTMENT OF
ENERGY

Office of
Science

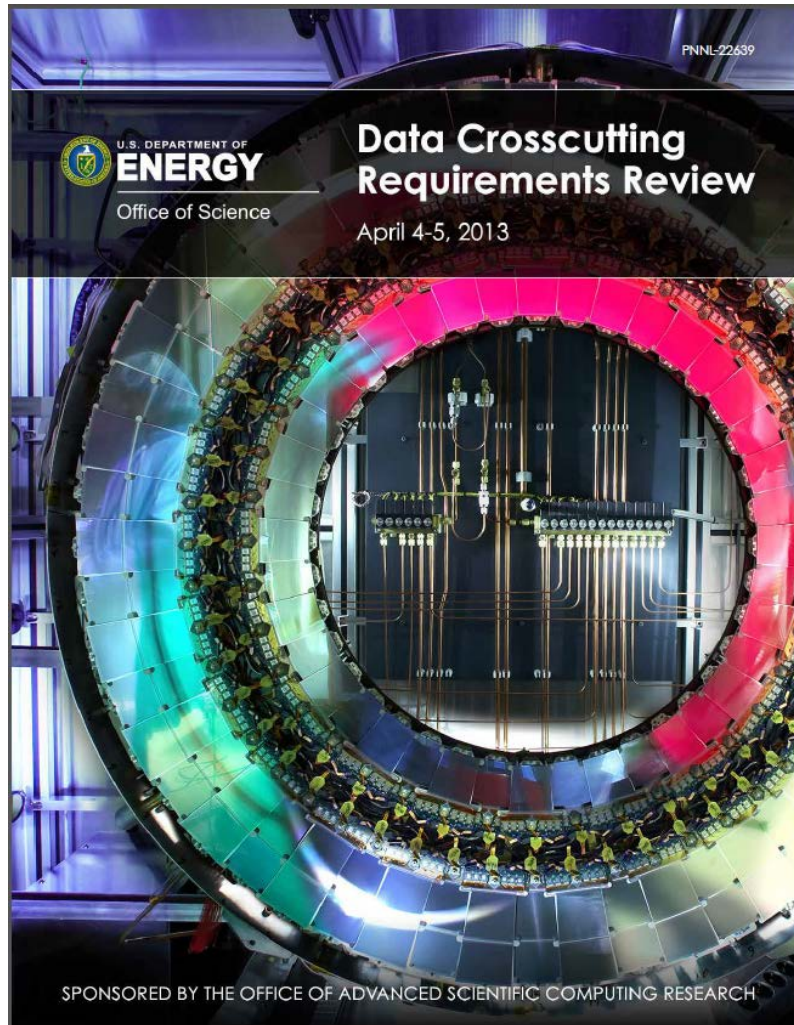
ASCR-BES Data Pilot Program

*...It is prudent for BES and ASCR to support some pilot studies that are targeted at specific scientific problems with strategic alignment for producing major breakthroughs. **Data and Communications in Basic Energy Sciences: Creating a Pathway for Scientific Discovery**
June 2012*

- **3 pilot projects were funded in FY2013**
- **Funding provided for 3 Post Docs to SciDAC Institute SDAV to collaborate with selected BES pilot projects**
 - A Pilot Collaboration to Apply Advanced Computing Capabilities to High Resolution Coherent Imaging of Energy Materials at Light Source Facilities: Develop real-time data pre-processing, reconstruction, and visualization codes for ptychography.
 - A Synchrotron Data Pilot: Develop a system that will enable users to perform data analysis and simulation on data intensive applications at the Advanced Light Source (ALS) using NERSC resources.
 - Advanced Structural Characterization using Experimental Scattering Data from Multiple Facilities: Develop a prototype of a federated data system providing central access to data, and to analysis tools, for selected beam lines and instruments at the Advanced Photon Source (APS) and at the Spallation Neutron Source (SNS).



Data Requirements Across Office of Science



HEP/ASCR Data Summit and Data Crosscutting Requirements Review

In April 2013, a diverse group of researchers from the U.S. Department of Energy (DOE) scientific community assembled in Germantown, Maryland to assess data requirements associated with DOE-sponsored scientific facilities and large-scale experiments.

http://science.energy.gov/~media/ascr/pdf/program-documents/docs/HEP_ASCR_Data_Summit_Report_April_2013.pdf

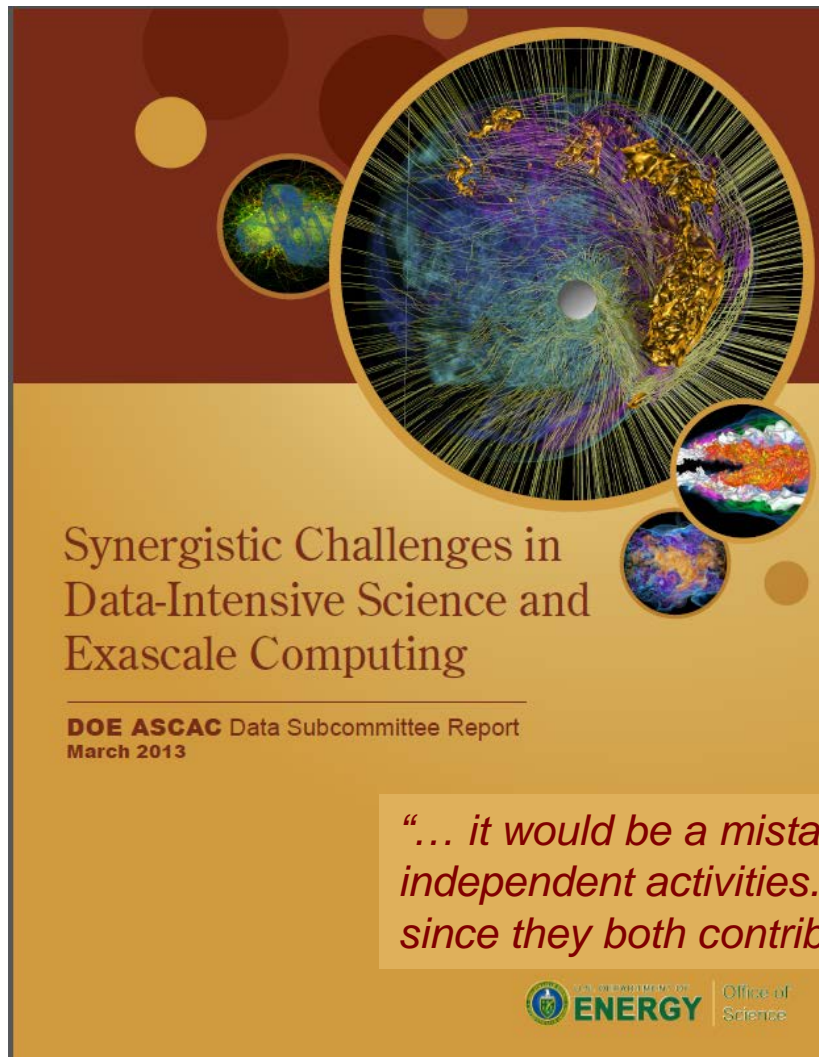
http://science.energy.gov/~media/ascr/pdf/program-documents/docs/ASCR_DataCrosscutting2_8_28_13.pdf



U.S. DEPARTMENT OF
ENERGY

Office of
Science

Exascale and Data Intensive Science



“... it would be a mistake to think of them [“big data” and “big compute”] as independent activities. Instead, their requirements are tightly intertwined since they both contribute to a shared goal of scientific discovery.”

- DOE missions require ASCR to address both simultaneously
- Data-intensive science faces many of the same technology challenges of exascale
 - Energy use is the grand challenge (e.g. the square kilometer array estimates 100MW needed for computing)
- **Advisory Committee charge**
 - Subcommittee looking at DOE mission needs, big data and exascale to identify synergies and high priority research needs

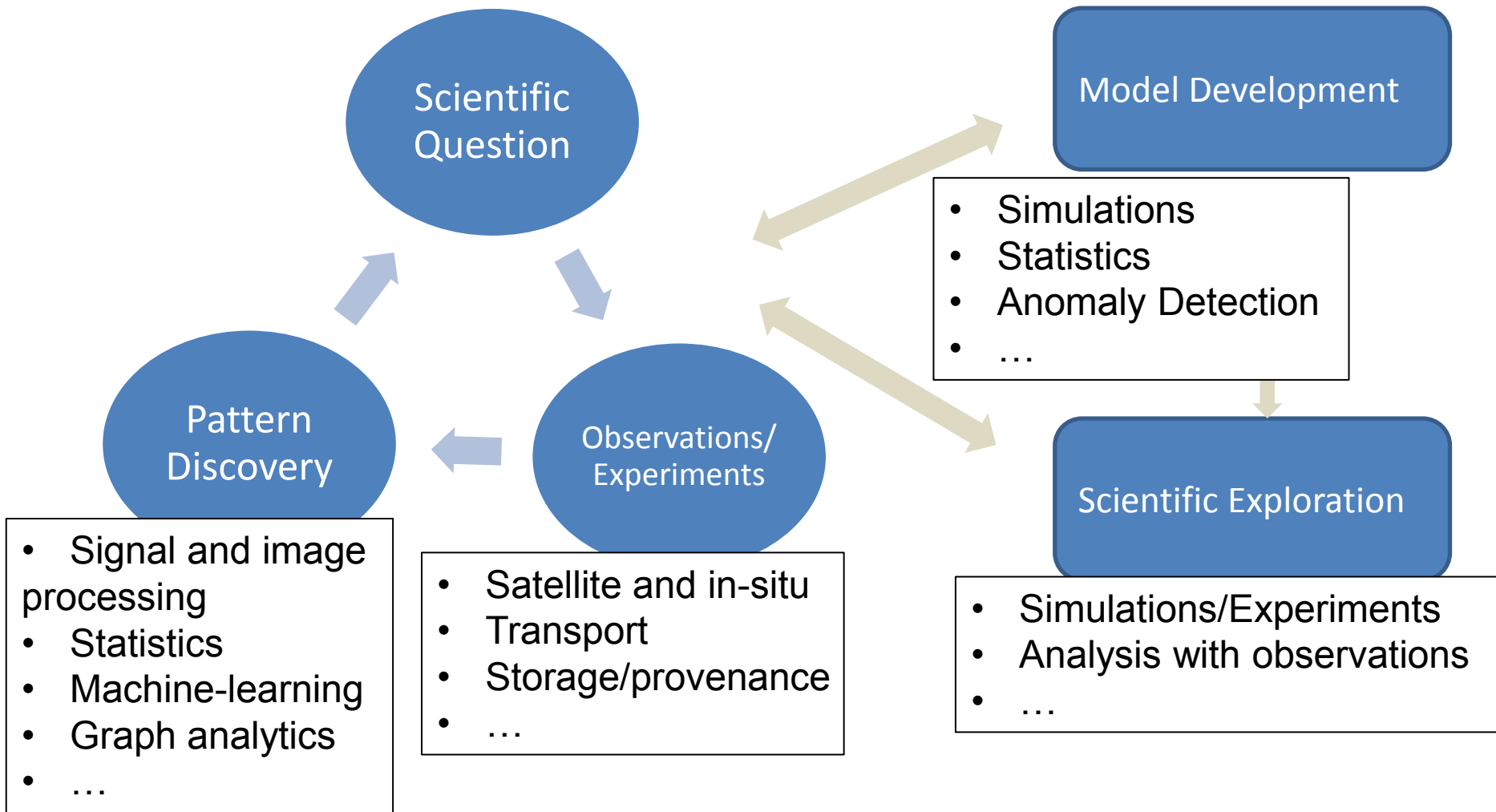
http://science.energy.gov/~media/ascr/ascac/pdf/reports/2013/ASCAC_Data_Intensive_Computing_report_final.pdf



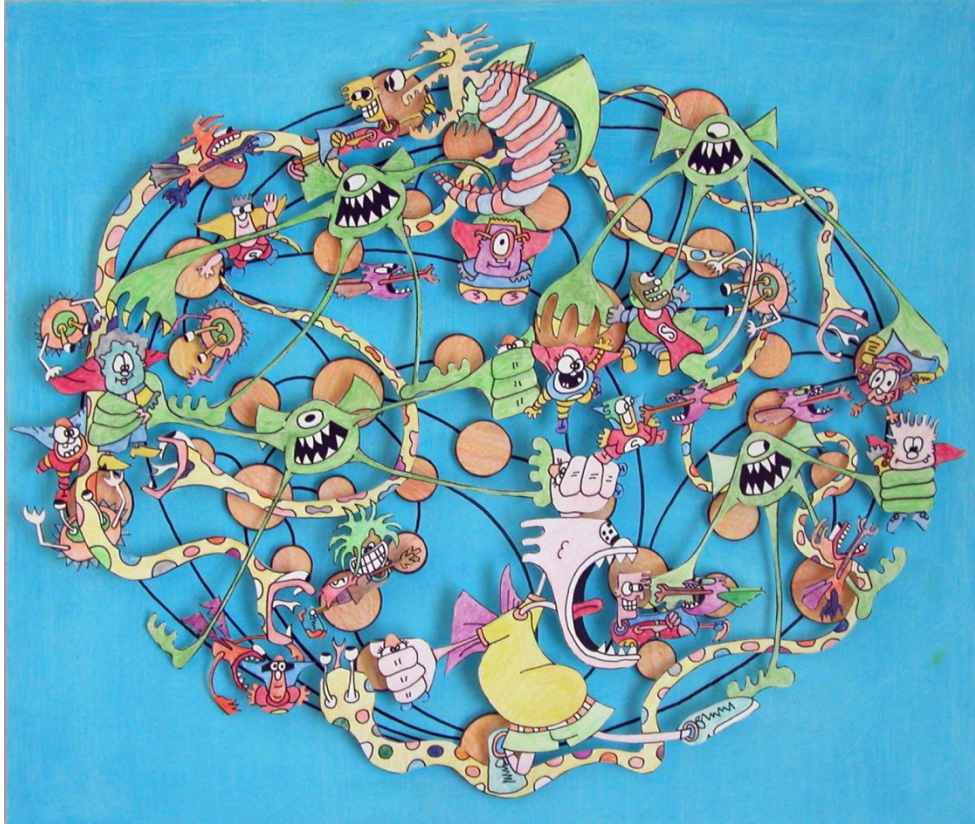
U.S. DEPARTMENT OF
ENERGY

Office of
Science

High Priority Investments for Data-Intensive Science



Our “Big Data”



THANK YOU!

Ceren Susut

Ceren.Susut-Bennett@science.doe.gov



U.S. DEPARTMENT OF
ENERGY

Office of
Science

