U.S. DEPARTMENT OF ENERGY | Office of Science

Big Data at the Department of Energy's Office of Science

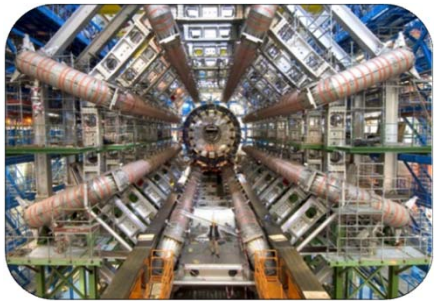
2nd NIST Big Data Public Working Group Workshop

June 1, 2017

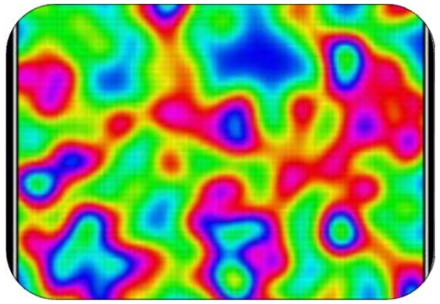
Laura Biven, PhD
Senior Science and Technology Advisor
Office of the Deputy Director for Science Programs
and Program Manager, Computer Science,
Advanced Scientific Computing Research
U.S. Department of Energy
Laura.Biven@science.doe.gov

Office of Science: Brief overview

Office of Science FY 2016: $5.35B

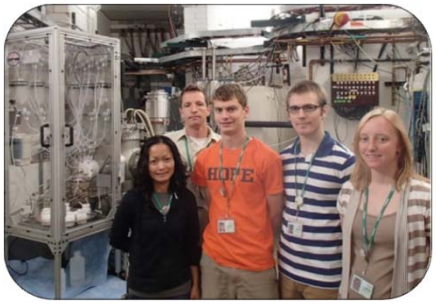


Largest Supporter of Physical Sciences in the U.S.*



Research: 42%, $2.2B

~40% of Research to Universities



> 22,000 Scientists Supported

Funding at >300 Institutions including all 17 DOE Labs

Construction:

13.5%, $723M

Facility Operations:

38%, $2.02B

>33,000 Scientific Facility Users**

* 43% of all physical sciences, 30% of computer science and math

** from all 50 states and DC

U.S. DEPARTMENT OF ENERGY | Office of Science

Slide from Dr. Murray

The Office of Science research portfolio

Advanced Scientific Computing Research	• **Delivering world leading computational and networking capabilities to extend the frontiers of science and technology**
Basic Energy Sciences	• **Understanding, predicting, and ultimately controlling matter and energy flow at the electronic, atomic, and molecular levels**
Biological and Environmental Research	• **Understanding complex biological, climatic, and environmental systems**
Fusion Energy Sciences	• **Building the scientific foundations for a fusion energy source**
High Energy Physics	• **Understanding how the universe works at its most fundamental level**
Nuclear Physics	• **Discovering, exploring, and understanding all forms of nuclear matter**

The DOE/SC Labs Today

Berkeley, California
202 acres and 97 buildings
3,396 FTEs
950 students & postdocs
9,320 facility users
www.lbl.gov

Richland, Washington
346 acres and 19 buildings
4,344 FTEs
550 students & postdocs
1,733 facility users
www.pnnl.gov

Ames, Iowa
8 acres and 12 buildings
308 FTEs
158 students & postdocs
www.ameslab.gov

Batavia, Illinois
6,800 acres and 354 buildings
1,720 FTEs
55 students & postdocs
2,097 facility users
www.fnal.gov

Argonne, Illinois
1,517 acres and 100 buildings
3,460 FTEs
1,054 students & postdocs
6,547 facility users
www.anl.gov

Menlo Park, California
426 acres and 151 buildings
1,596 FTEs
213 students & postdocs
4,474 facility users
www.slac.stanford.edu

Oak Ridge, Tennessee
4,421 acres and 194 buildings
4,586 FTEs
1,080 students & postdocs
3,215 facility users
www.ornl.gov

Newport News, Virginia
169 acres and 72 buildings
729 FTEs
60 students & postdocs
1,261 facility users
www.jlab.org

Princeton, New Jersey
89 acres and 34 buildings
429 FTEs
54 students & postdocs
290 facility users
www.pppl.gov

Upton, New York
5,322 acres and 310 buildings
2,882 FTEs
642 students & postdocs
4,134 facility users
www.bnl.gov

Office of Science User Facilities

28 world-leading facilities serving over 33,000 researchers annually

- **supercomputers,**
- **high intensity x-ray, neutron, and electron sources,**
- **nanoscience facilities,**
- **genomic sequencing facilities,**
- **particle accelerators,**
- **fusion/plasma physics facilities, and**
- **atmospheric monitoring capabilities.**

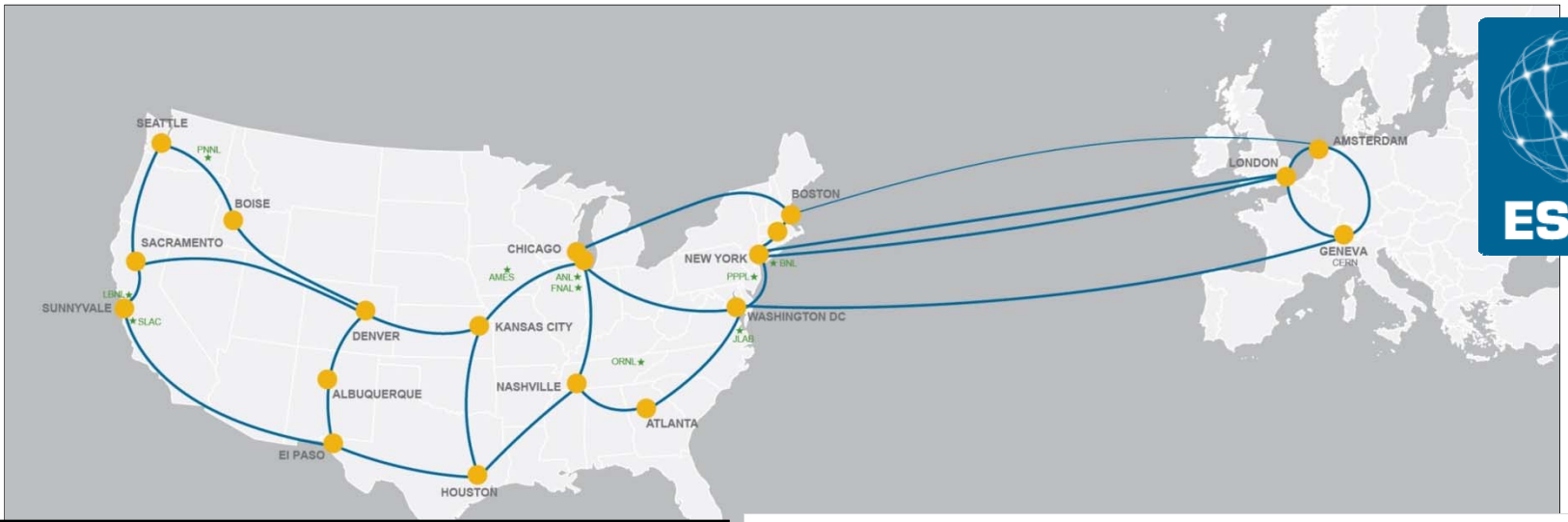
- **Open access; allocation determined through peer review of proposals**

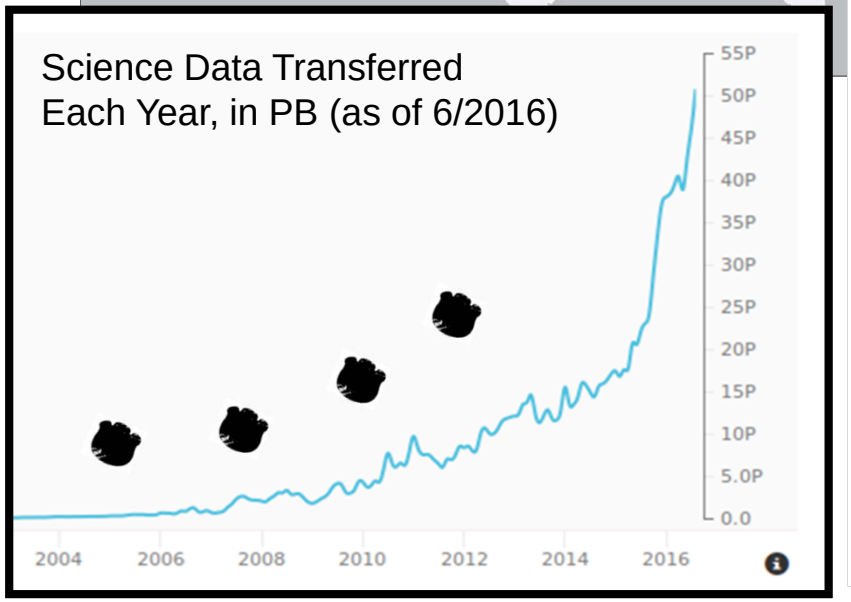
- **Free for non-proprietary work published in the open literature**

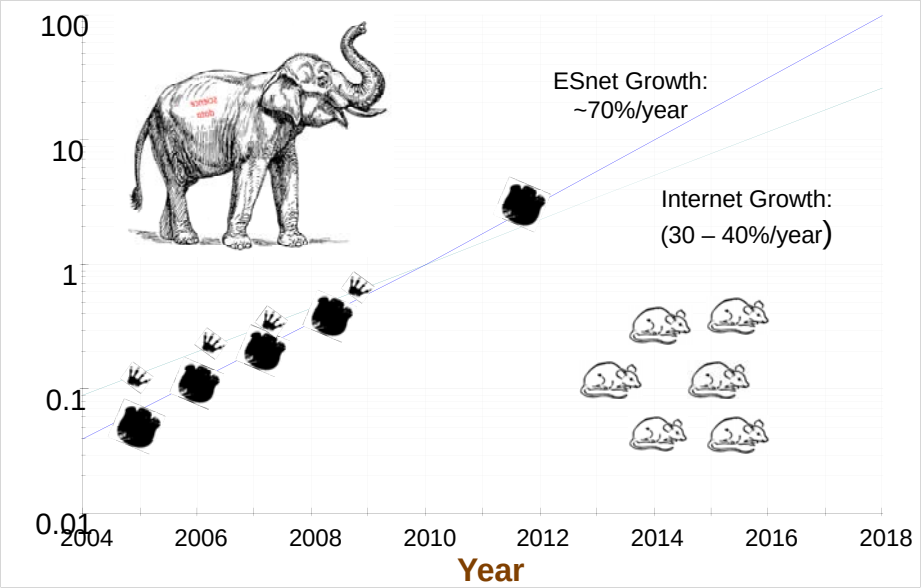
- Full cost recovery for proprietary work

Signatures of Big Data Challenges

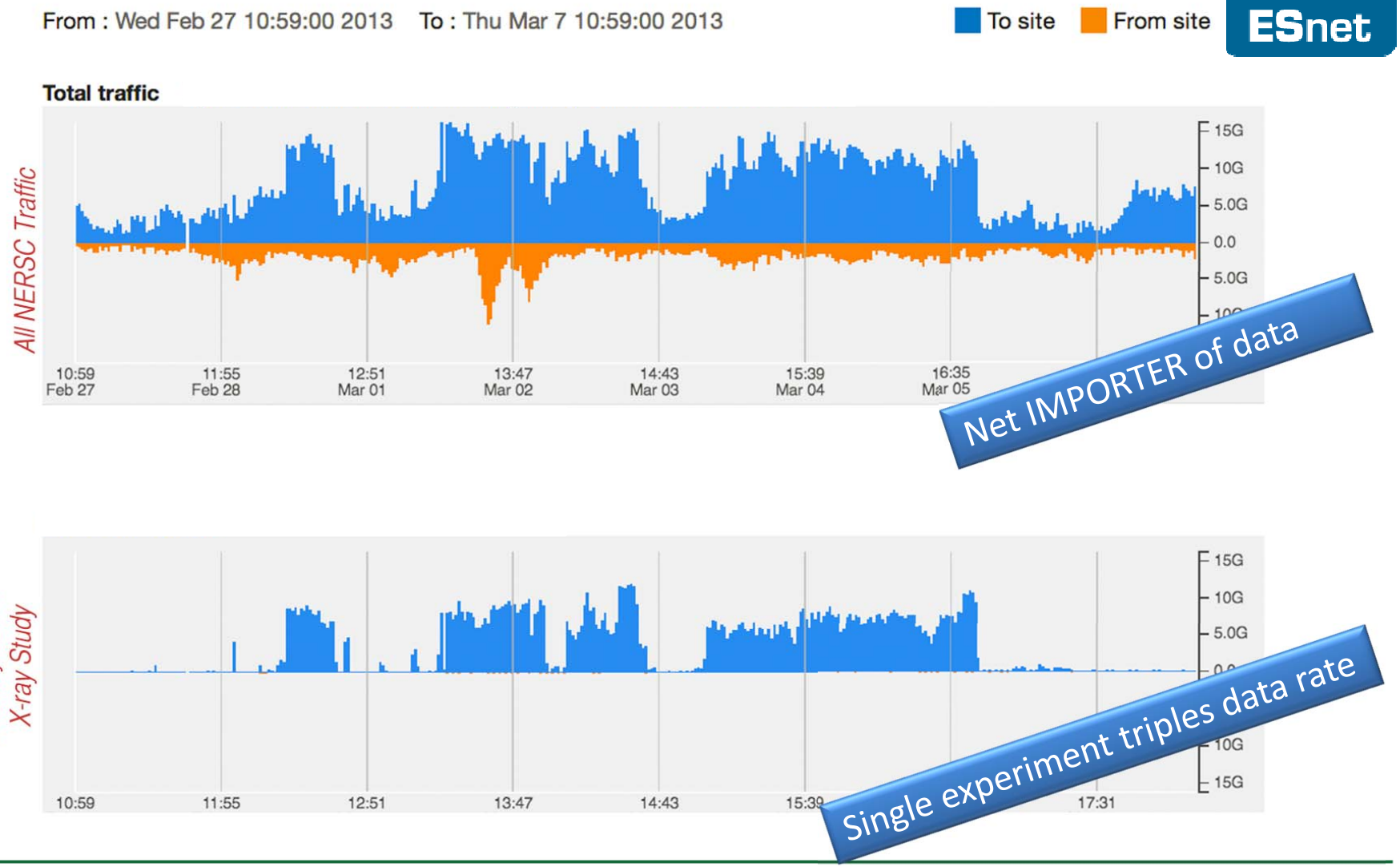
Data-Intensive Science Drives Exponential Network Growth





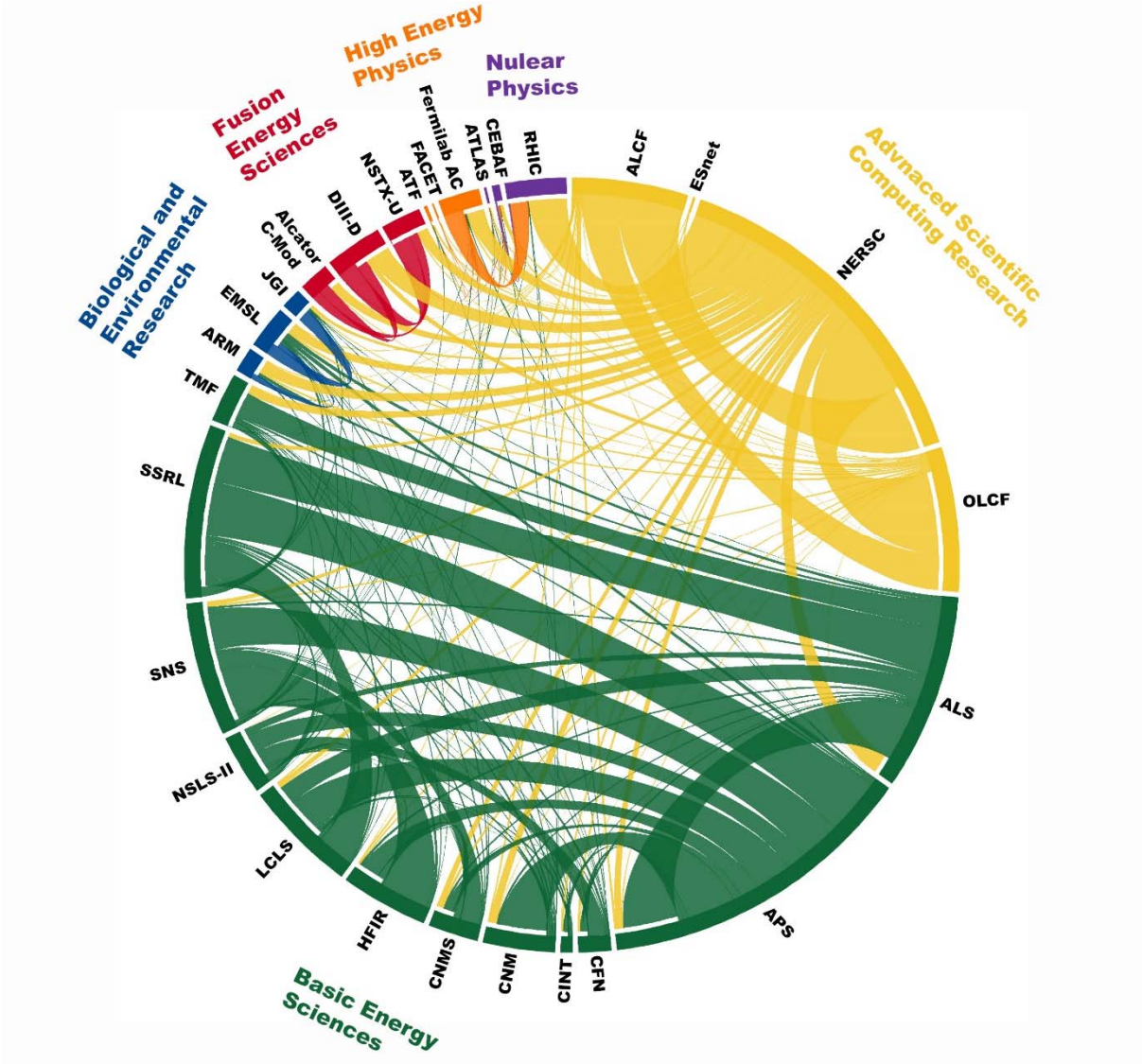


Single Light Source Data Flow Triples NERSC's Network Usage



User Crossover Among SC User Facilities by Program, FY 2016

The width of the ribbon connecting two facilities corresponds to the number of users who utilized both of those facilities



Over **3,200** users performed research at two or more facilities in FY2016

Big Data,
Big Changes in Science

Advanced Light Source

Mission focus:
Synchrotron light source for imaging, scattering, and spectroscopy experiments in chemical, geological, life, material and physical sciences.
Users come, do experiment, want to leave with data in hand.

"...it is getting to the point where users cannot just download their data: their hard drive isn't big enough, and if it was, they wouldn't have adequate computing power to do anything with it."

Impediments/Gaps

- Increasing use of simulation to design experiments
- Increasing use of simulation to design science user facilities
- Changing science paradigm: Data too large for users to take home to analyze
- Data volume, rate exceed capacity and capability – need to use HPC to analyze data near-real time
- Desire to steer experiment's data collection
- Diversity of science: one accelerator, O(40) beamlines, diverse experiments at each: no "one size fits all"
- New data vis/analysis methods
- Challenge: Usability & accessibility of computing, data

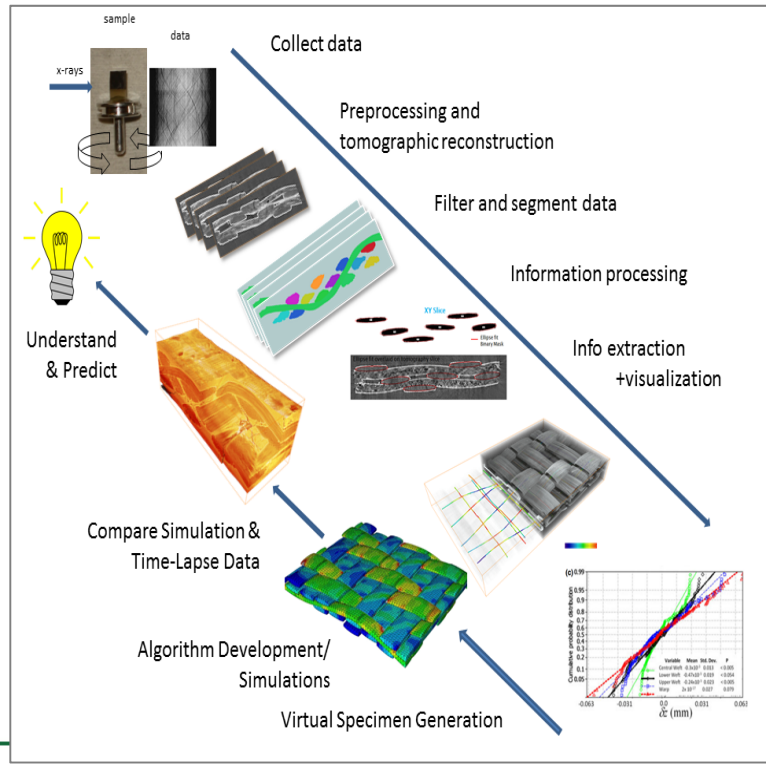


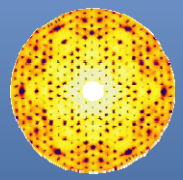
Image: Dula Parkinson (LBNL)

Data Driven Neutron Scientific Discovery Enabled by HPC

Data Base Generation

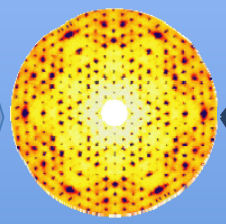
Neutron modeling

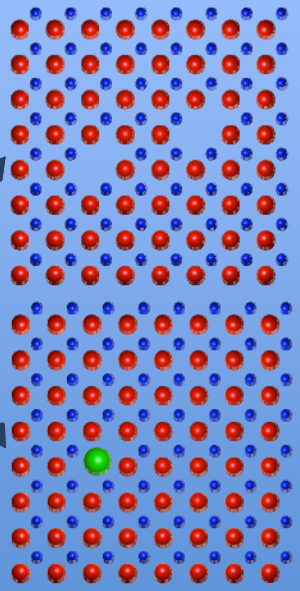
Experimental Neutron Data



Utilize both computational and experimental production of diffuse scattering data

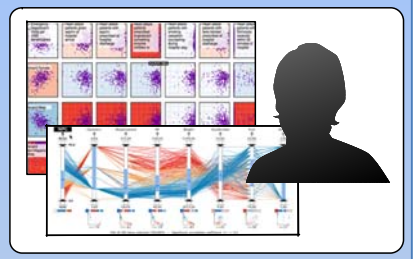
Analysis and Learning





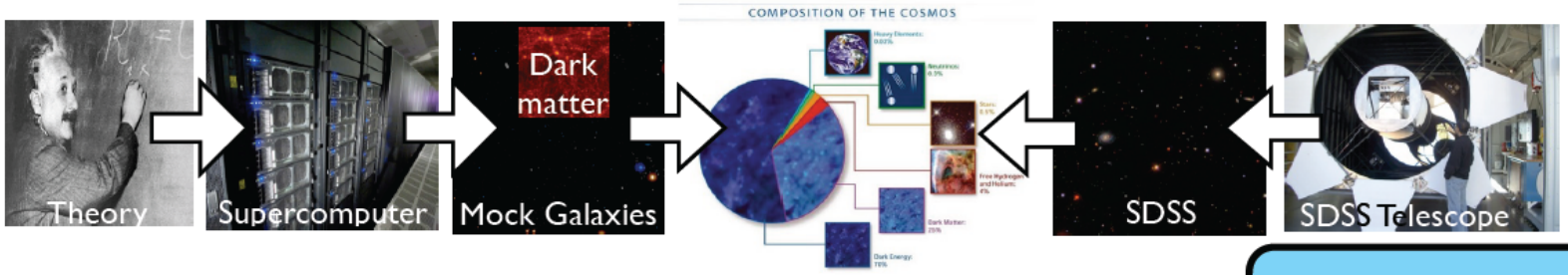
Computational methods can be generated to identify material properties and anomalies from massive & expanding database

Scientific Discovery

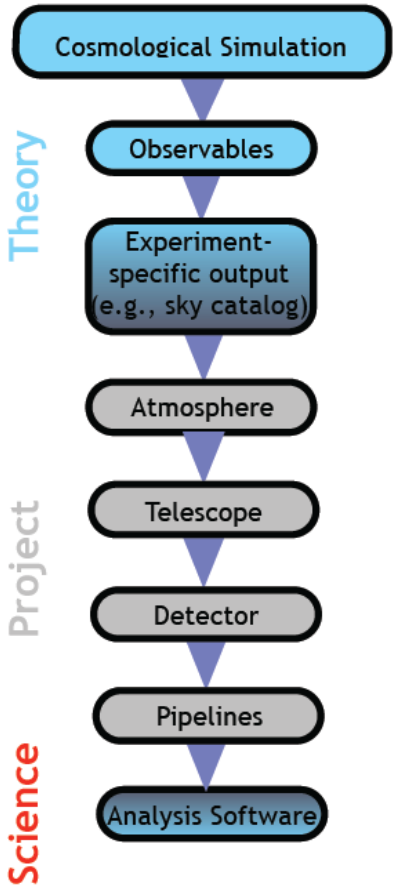


Scientific analysis and search tools providing access to massive database of diffuse scattering knowledge

Computational Cosmology: Role of Simulations



- **Three Roles of Cosmological Simulations**
  - Basic theory of cosmological probes
  - Production of high-fidelity 'mock skies' for end-to-end tests of the observation/analysis chain
  - Essential component of analysis toolkits: Control systematics
- **Extreme Simulation and Analysis Challenges**
  - Large dynamic range simulations; control of subgrid modeling and feedback mechanisms
  - Design and implementation of complex analyses on large datasets; new fast (approximate) algorithms
  - Solution of large statistical inverse problems of scientific inference (many parameters, ~10-100) at the ~1% level



Convergence of Data and Compute

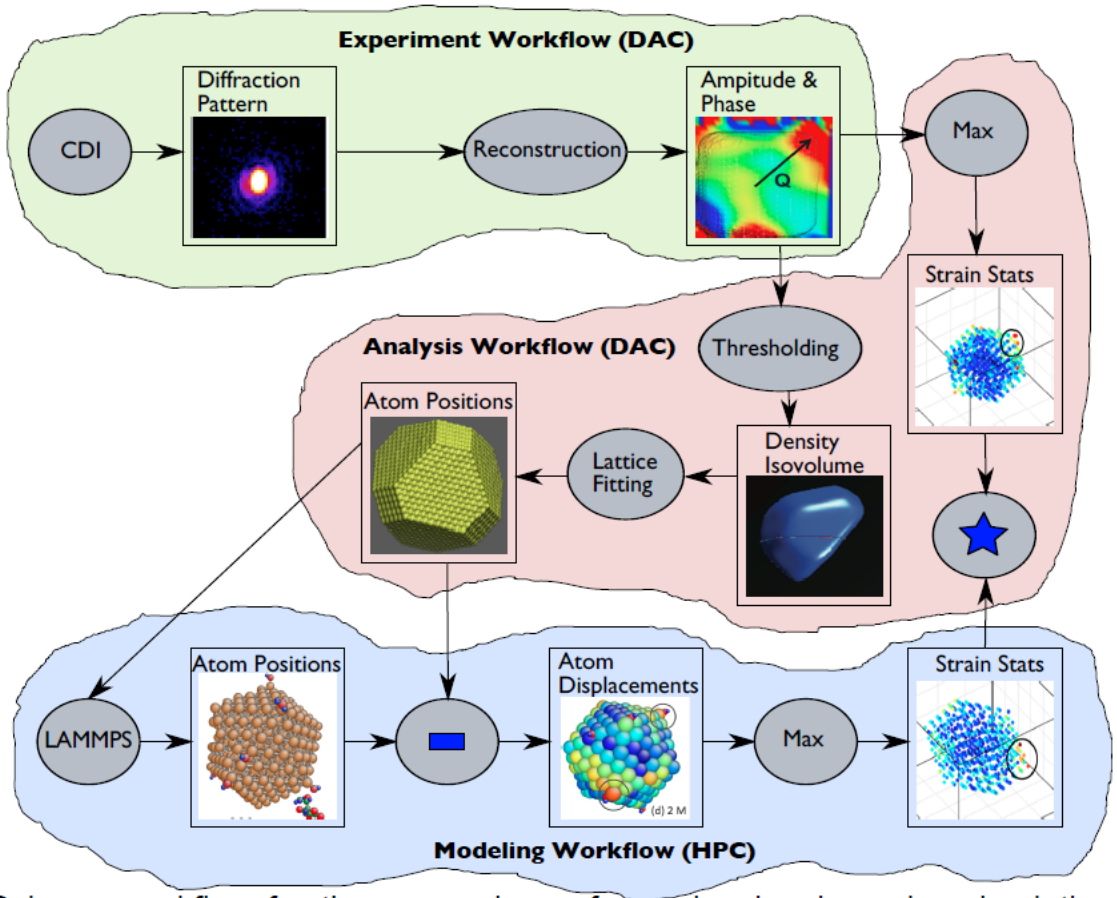
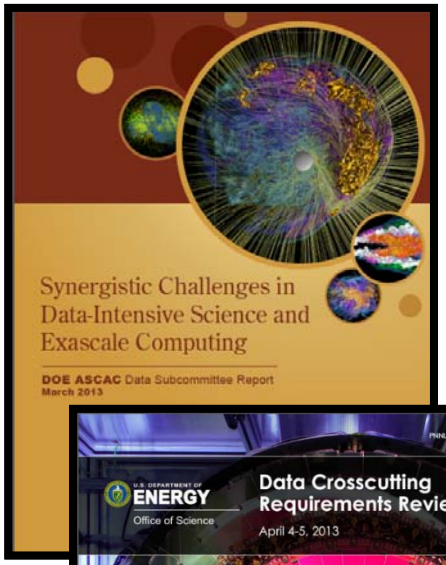


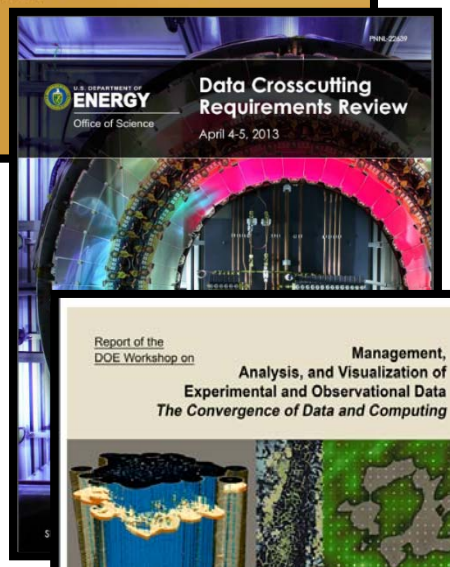
Figure 1: Science workflow for the comparison of a molecular dynamics simulation with a high-energy X-ray microscopy of the same material system includes three interrelated computational and experimental workflows.

Image from "The Future of Scientific Workflows"
https://science.energy.gov/~/media/ascr/pdf/programdocuments/docs/workflows_final_report.pdf

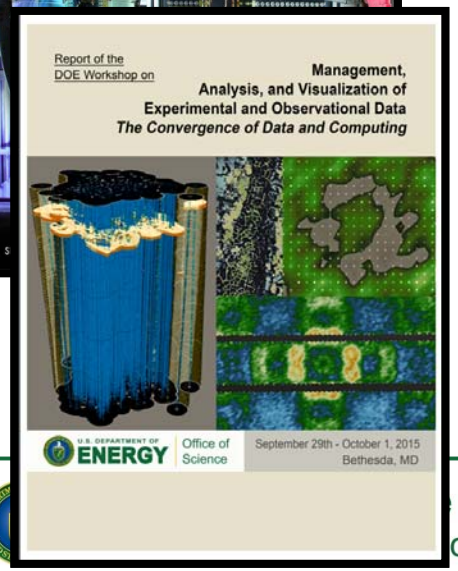
Reports



Synergistic Challenges in Data-Intensive Science and Exascale Computing (ASCAC Report)
This new report discusses the natural synergies among the challenges facing data-intensive science and exascale computing, including the need for a new scientific workflow.
http://science.energy.gov/~/media/ascr/ascac/pdf/reports/2013/ASCAC_Data_Intensive_Computing_report_final.pdf



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/ASCR_DataCrosscutting2_8_28_13.pdf



Management, Analysis, and Visualization of Experimental and Observational Data: The Convergence of Data and Computing
The purpose of this workshop…is to help the Advanced Scientific Computing Research (ASCR) and research community better understand needs related to the management, analysis, and visualization of experimental and observational data (EOD) collected and generated by experimental and observational science projects (EOS) at Office of Science user facilities.
http://science.energy.gov/~/media/ascr/pdf/programdocuments/docs/ascr-eod-workshop-2015-report_160524.pdf

Reports

The Future of Scientific Workflows (ASCR Workshop Report)
The mission of this workshop was to develop requirements for workflow methods and tools in a combined high-performance computing (HPC) and distributed-area instruments and computing (DAIC) work environment, in order to enable science applications to better manage their end-to-end data flow.
https://science.energy.gov/~/media/ascr/pdf/programdocuments/docs/workflows_final_report.pdf

STREAM2016: Streaming Requirements, Experience, Applications and Middleware Workshop
This report describes the discussions, outcomes, and conclusions from STREAM2016: Streaming Requirements, Experience, Applications and Middleware Workshop – the second workshop in the STREAM series, held on March 22-23, 2016 in Tysons, VA. STREAM2016 focused on DOE applications, computational and experimental facilities, as well as software systems. The role of streaming and steering as a critical aspect of the linkage between experimental and computing facilities was pervasive throughout the workshop. Given the overlap in interests and challenges faced by industry, the workshop had significant presence from several major companies in this area.
https://science.energy.gov/~/media/ascr/pdf/programdocuments/docs/2017/STREAM2016.pdf

Computing and Data Requirements in the Exascale Age
The DOE Office of Science brought together key computational domain scientists, and DOE planners and administrators to determine the requirements for an exascale ecosystem that includes computation, data analysis, software, workflows, HPC services, and the full-scale range of computer requirements needed to support forefront scientific research through 2025.
http://exascaleage.org/

Summary: Challenges

- Our ability to collect data far exceeds our ability to analyze and store data
- Increasing difficulty, inability to gain knowledge from data
- Critical streaming and steering needs at Scientific User Facilities
- Workflows are becoming more complex, relying on distributed resources and multi-modal data
- Convergence of data and computing: data- and computing-centric needs increasingly intertwined, symbiotic