U.S. DEPARTMENT OF ENERGY | Office of Science

Research Data Management

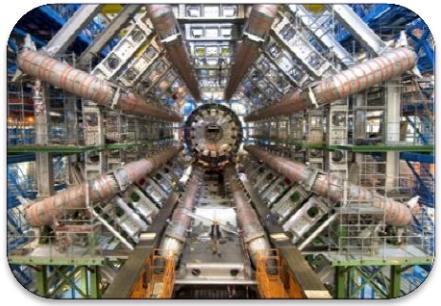
NITRD FASTER

July 18, 2016

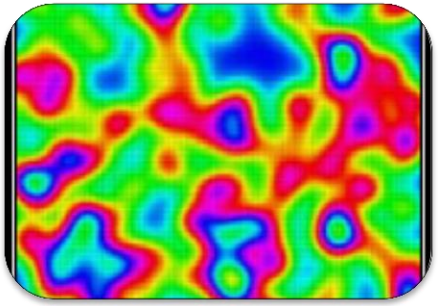
Laura Biven, PhD
Senior Science and Technology Advisor
Office of the Deputy Director for Science Programs (SC-2)
U.S. Department of Energy
Laura.Biven@science.doe.gov

Office of Science FY 2016: $5.35B

Dr. Murray slide

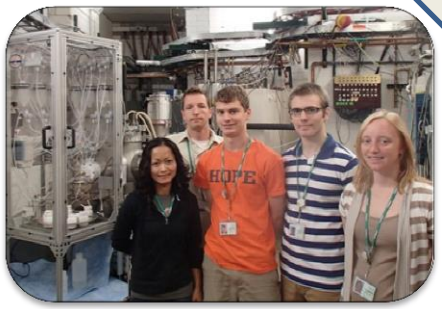


Largest Supporter of Physical Sciences in the U.S.*



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

The Office of Science research portfolio

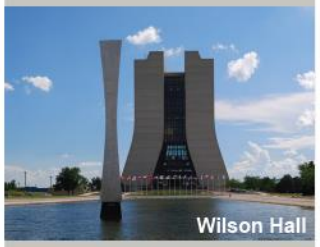
Advanced Scientific Computing Research	• **Delivering world leading computational and networking capabilities to extend the frontiers of science and technology**
Basic Energy Sciences	• **Understanding, predicting, and ultimately controlling matter and energy flow at the electronic, atomic, and molecular 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

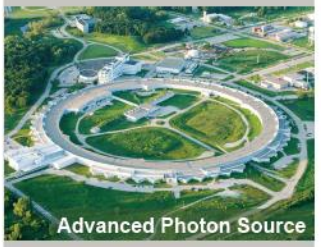
Richland, Washington
346 acres and 19 buildings
4,344 FTEs
550 students & postdocs
1,733 facility users
www.pnnl.gov

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



Wilson Hall

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



Advanced Photon Source

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

Spallation Neutron Source

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

NSTX Spherical Tokamak

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

Relativistic Heavy Ion Collider

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

FY 2015 User Crossover Analysis

	C-Mod	ALCF	ALS	APS	ARM	ATF	ATLAS	CEBAF	CFN	CINT	CNM	CNMS	DIII-D	EMSL	Esnet	FACET	Fermilab AC	HFIR	JGI	LCLS	NERSC	NSLS-II	NSTX-U	OLCF	RHIC	SNS	SSRL	TMF
Alcator C-Mod	**224**	2	1										48								40		35	2				
ALCF	2	**990**	4	18	9		1	5	2	3	15		3	12	1	1	10	3		4	281		2	169	2	4	3	2
ALS	1	4	**2560**	286	13			6	6	9	8	4		18			2	13	2	65	90	5			1	15	223	96
APS		18	286	**5471**			8	14	19	8	154	29		19		2	5	77	3	77	26	25	1	1		163	183	5
ARM		9	13		**1121**			2						23			2				71			15			3	
ATF						**75**		1								10					4				2			
ATLAS		1		8			**392**	5		1	5										3			1				
CEBAF		5	6	14	2	1	5	**1510**	1	1	7	3		3			36	2		1	44			14	29	7	4	1
CFN		2	6	19				1	**493**	1	2	3		1				3		1	19	12		2		3	3	2
CINT		3	9	8			1	1	1	**502**		9						2			6	1		4		2	8	2
CNM		15	8	154			5	7	2		**529**			2			4	2		4	16	2			1		8	1
CNMS			4	29				3	3	9		**575**				1	1	29	1		16			3		74	1	2
DIII-D	48	3											**557**				1				86		87	14				1
EMSL		12	18	19	23			3	1		2			**713**			2	2	39		64			13		3	12	3
Esnet		1													**48**						2							
FACET		1		2		10						1				**148**	1			3	10					1	3	
Fermilab AC		10	2	5	2			36			4	1	1	2		1	**1924**		2	1	45		3	4	81	3	2	2
HFIR		3	13	77				2	3	2	2	29		2				**492**	2	2	4			3		235	7	
JGI			2	3								1		39			2	2	**957**	1	8					1	2	
LCLS		4	65	77				1	1		4					3	1	2	1	**829**	14	3		1		1	93	
NERSC	40	281	90	26	71	4	3	44	19	6	16	16	86	64	2	10	45	4	8	14	**6332**	2	41	304	142	13	14	30
NSLS II			5	25					12	1	2									3	2	**95**				1	6	1
NSTX-U	35	2		1									87				3				41		**356**	4				
OLCF	2	169		1	15		1	14	2	4		3	14	13			4	3		1	304		4	**1107**	1	4		
RHIC		2	1			2		29			1						81				142			1	**1015**	1		
SNS		4	15	163				7	3	2		74		3		1	3	235	1	1	13	1		4	1	**843**	7	
SSRL		3	223	183	3			4	3	8	8	1		12		3	2	7	2	93	14	6				7	**1626**	31
TMF		2	96	5				1	2	2	1	2	1	3			2				30	1					31	**677**
Total Crossover Users	128	556	867	1123	138	17	24	186	80	57	231	176	240	216	3	32	207	386	61	271	1395	58	173	559	260	538	613	179
Crossover %	57%	56%	34%	21%	12%	23%	6%	12%	16%	11%	44%	31%	43%	30%	6%	22%	11%	78%	6%	33%	22%	61%	49%	50%	26%	64%	38%	26%

Office of Science and Big Data Growth

http://science.energy.gov/~/media/ascr/pdf/programdocuments/docs/ascr-eod-workshop-2015-report_160524.pdf

	Current/ Near Term		Long Term	
Use Case	**Max rate**	**Monthly/Annual Totals**	**Max Rate**	**Monthly/Annual Totals**
EMSL	100 Mbps	3.6 TB monthly	1 GB/s	18 TB monthly
Climate Simulation	5 TB/day	1.5 PB annually	250 TB/day	50 PB annually
ARM		18 TB monthly		5 PB obs data & 1 PB model data annually
ALS	10 Gbps	140 TB monthly	80 Gbps	1.5 PB monthly
LCLS	1-10 Gbps	1.5 PB annually	100 Gbps	15 PB annually
ORNL Neutron sources	500 MB/s	.3 PB annually	5 GB/s	1 PB annually
Scanning probe and electron microscopies	10-100 GB/day		10 Mb/s	
APS	5 TB/day	2 PB annually	1 PB/day	500 – 1000 PB annually
DUNE	10 GB/s	500-800 TB annually	100 GB/s	10-20 PB annually

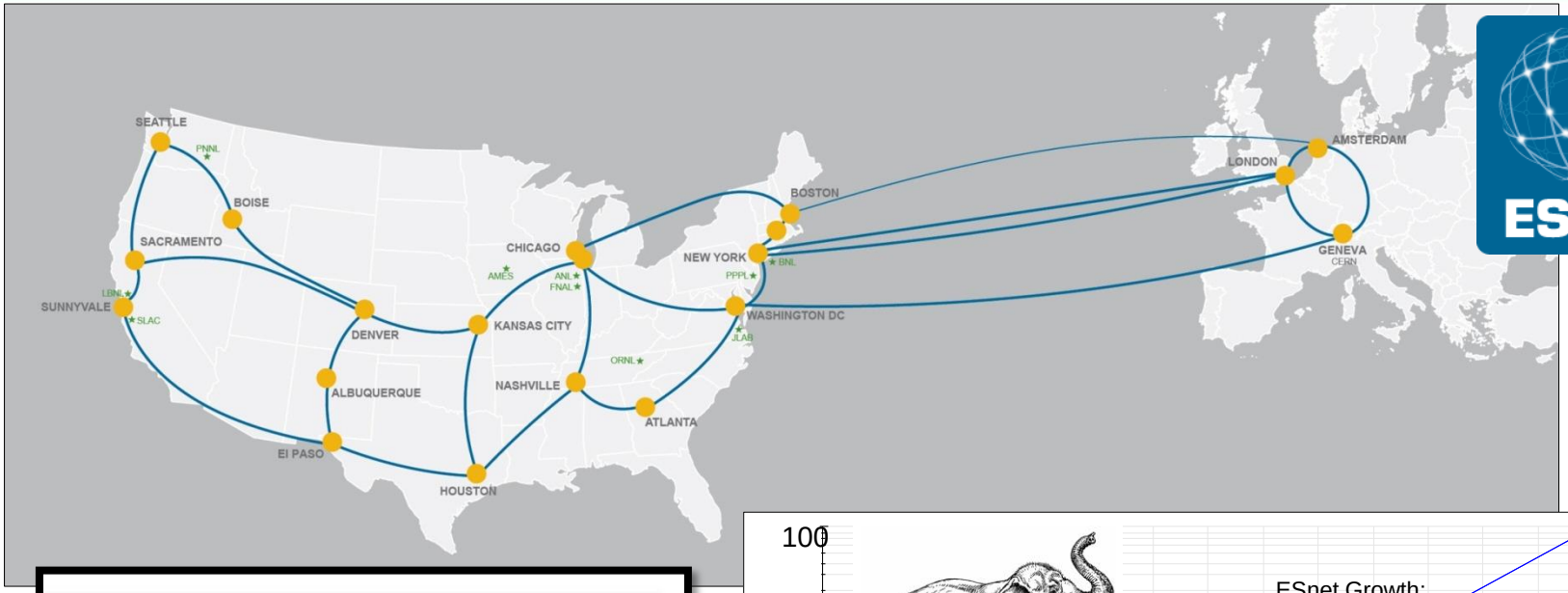
Big Data, Big Compute, Big Networks

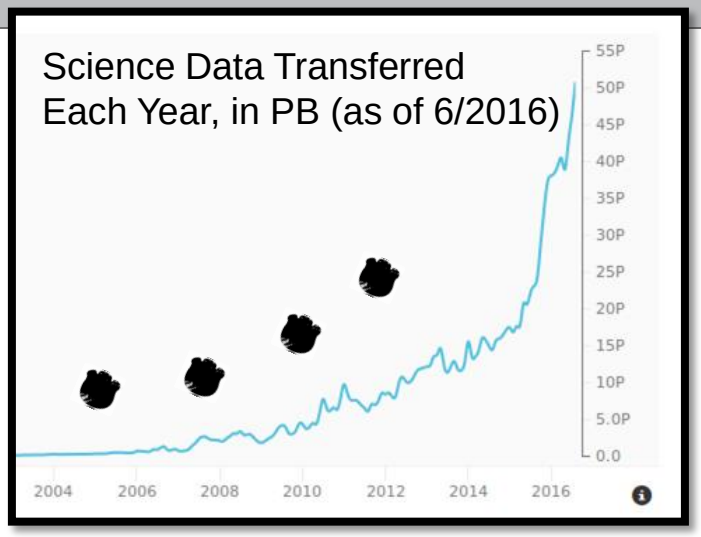
- "… trend toward the convergence of data and computing."
  Management, Analysis, and Visualization of Experimental and Observational Data: The Convergence of Data and Computing

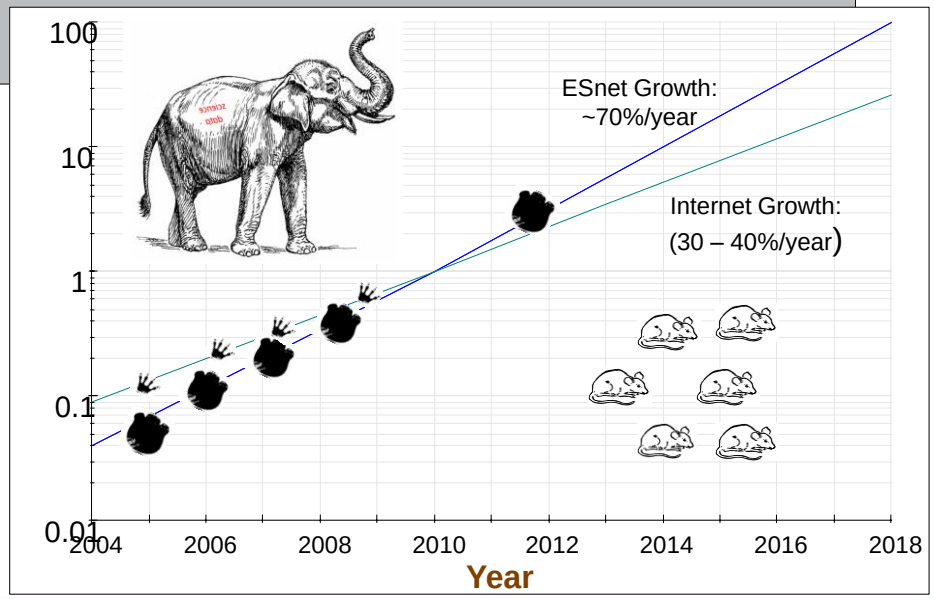
- "… it would be a mistake to think of them ["big data" and "big compute"] as independent activities. Instead, their requirements are tightly intertwined since they both contribute to a shared goal of scientific discovery."
  ASCAC report on Synergistic Challenges in Data-Intensive Science and Exascale Computing

- For many problems there is a deep coupling of observation (measurement) and computation (simulation)

Data-Intensive Science Drives Exponential Network Growth

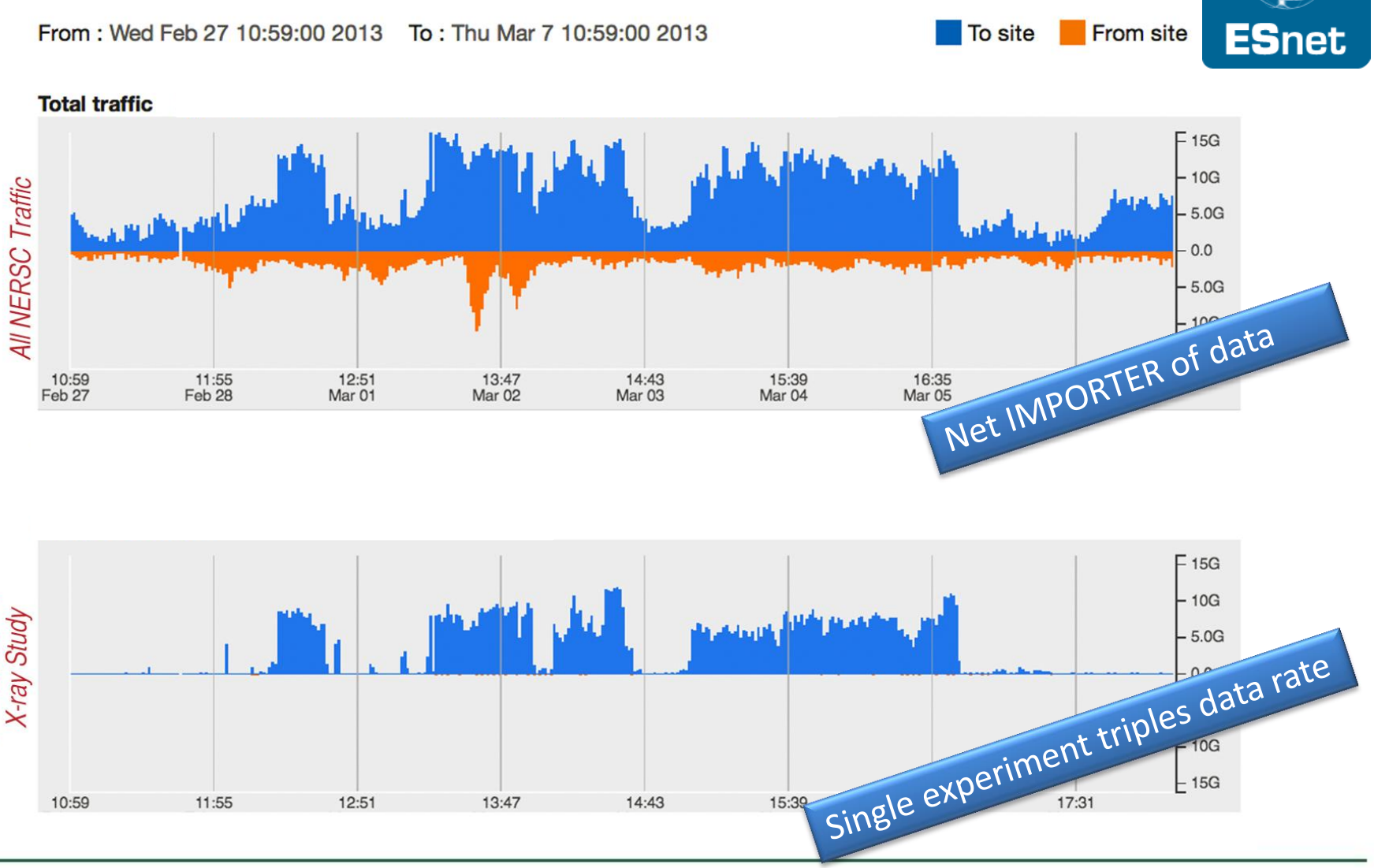






U.S. DEPARTMENT OF ENERGY | Office of Science

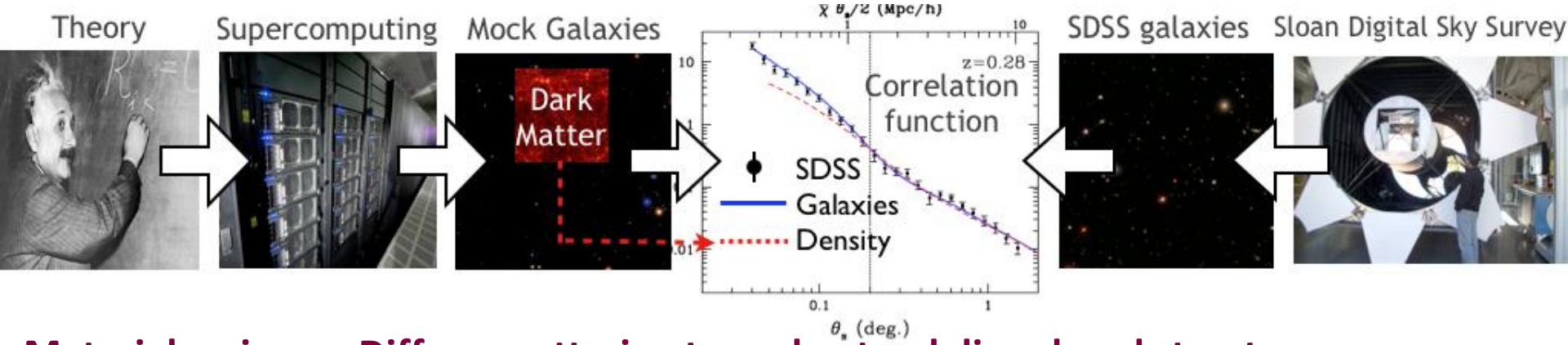
Single Light Source Data Flow Triples NERSC's Network Usage



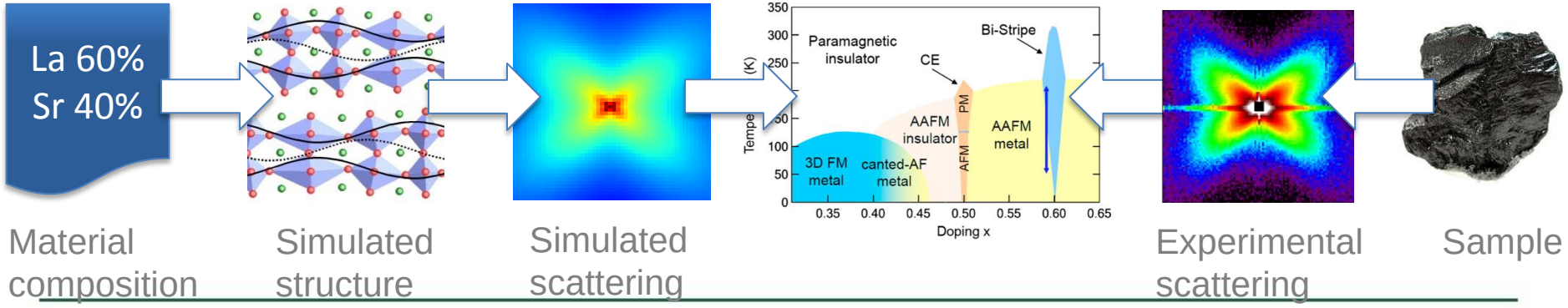
Data-Driven Science Examples

For many problems there is a deep coupling of observation (measurement) and computation (simulation)

Example: Cosmology – the study of the universe as a dynamical system

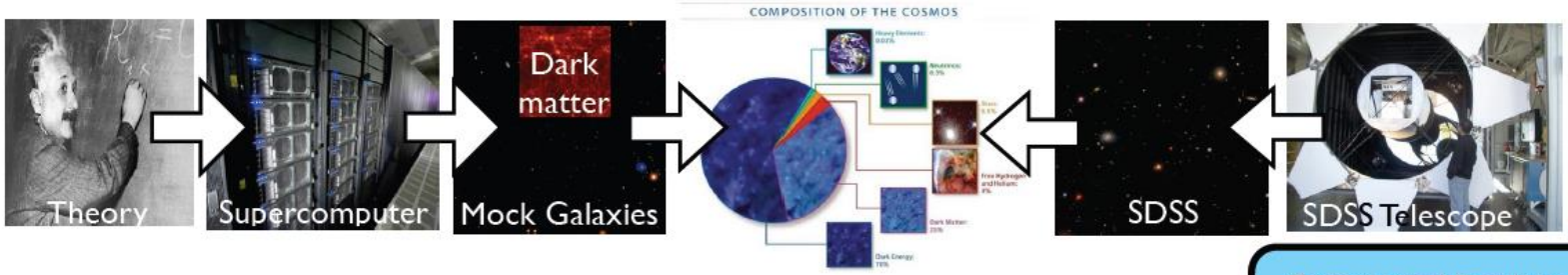


Materials science: Diffuse scattering to understand disordered structures

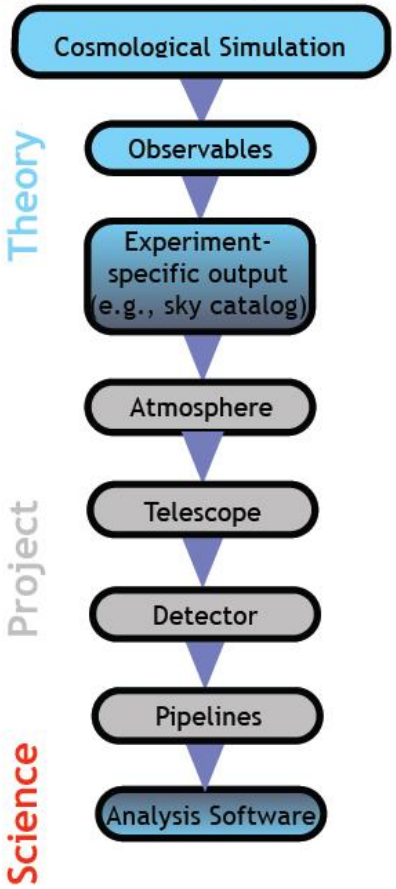


Images from Salman Habib et al. (HEP, MCS, etc.) and Ray Osborne et al. (MSD, APS, etc.)

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



Data Driven Neutron Scientific Discovery Enabled by HPC

Data Base Generation

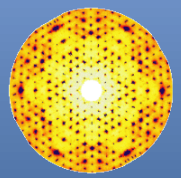
Neutron modeling

DISCUS SIMULATION PACKAGE

DANSE

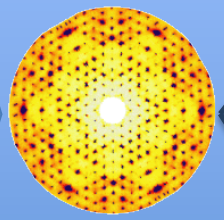
Experimental Neutron Data

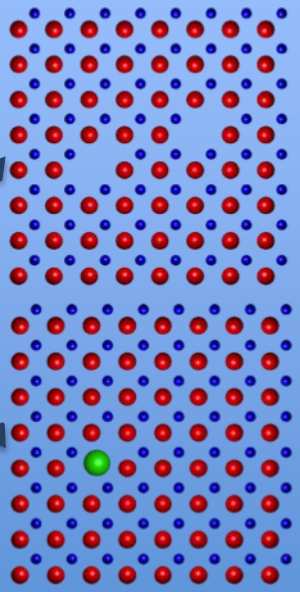
SNS SPALLATION NEUTRON SOURCE



Utilize both computational and experimental production of diffuse scattering data

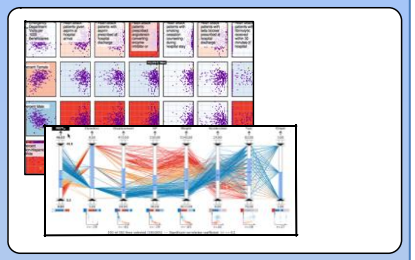
Analysis and Learning





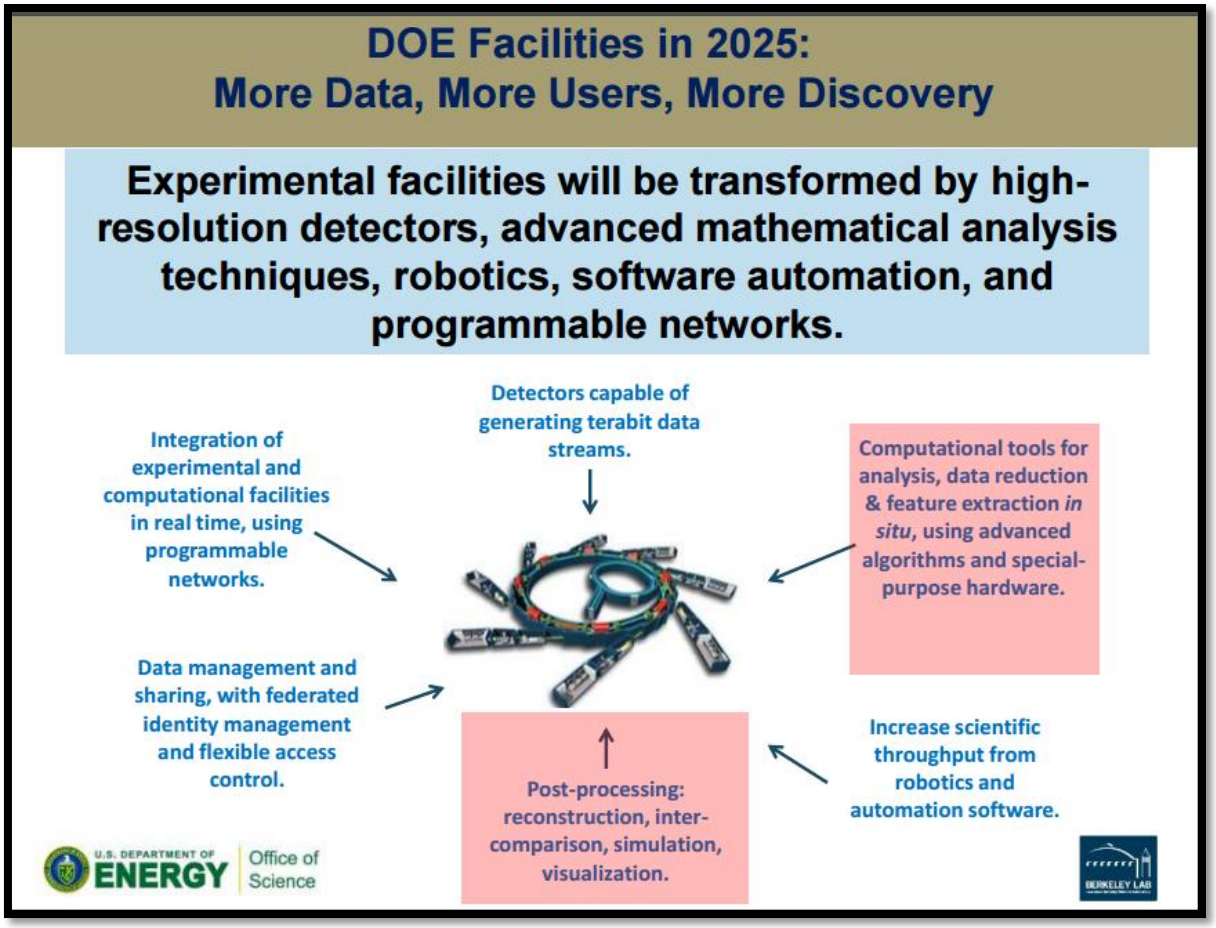
Computational methods can be generated to identify material properties and anomalies from massive & expanding database

Scientific Discovery



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

More Data, More Users, More Discovery



Applied Mathematics for Experimental Science, J.Sethian, ASCAC July, 2015
http://science.energy.gov/~/media/ascr/ascac/pdf/meetings/20150727/Sethian_ascac_july_2015.pdf

Atmospheric Radiation Measurement

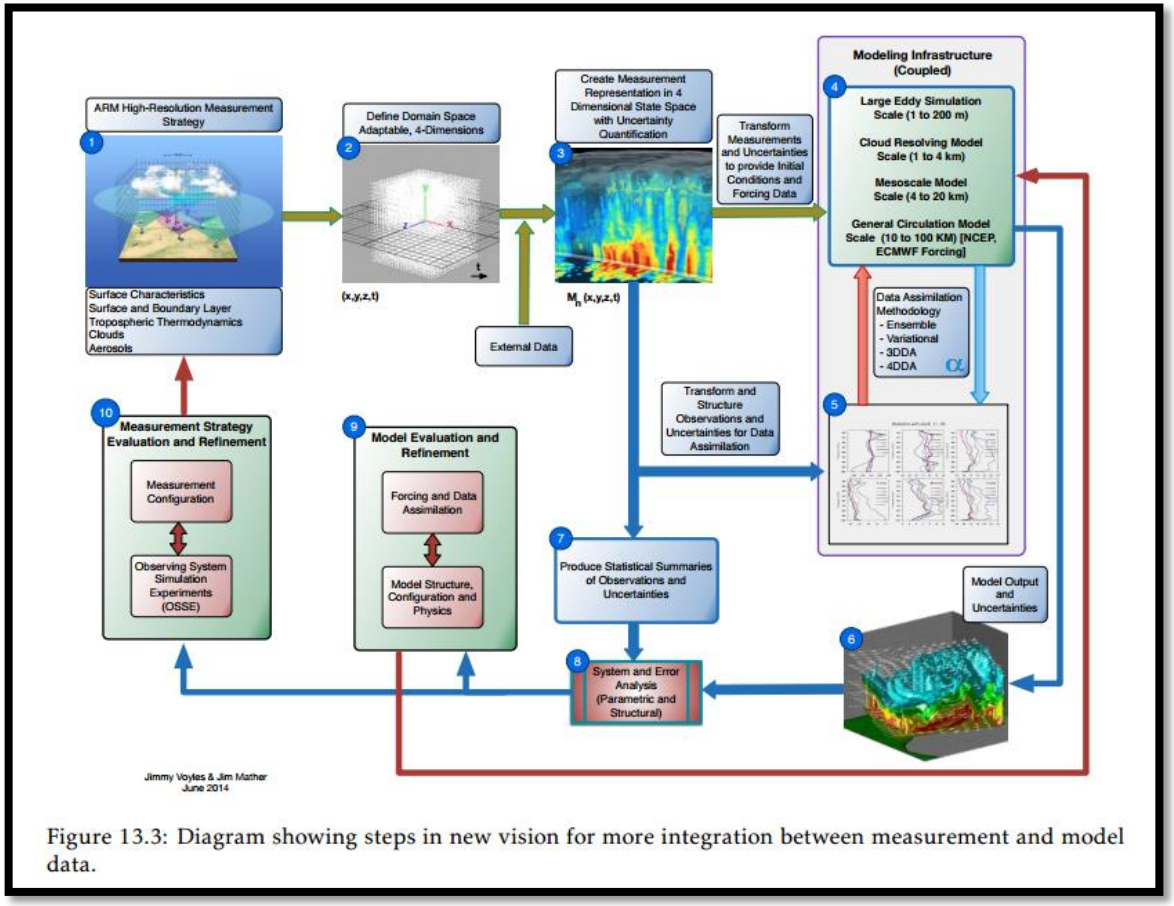
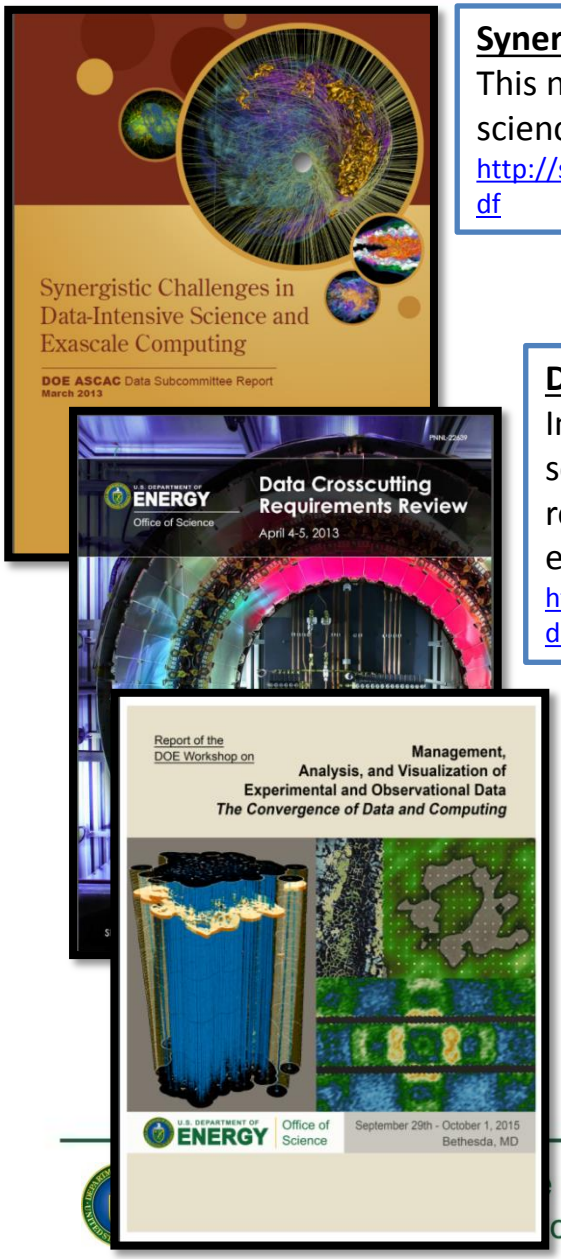


Figure 13.3: Diagram showing steps in new vision for more integration between measurement and model data.

ARM is now beginning a new approach where high resolution Large Eddy Scale (LES) models forced by observational data will be run routinely in order to fill in some of the gaps inherent in observational dataset.

http://science.energy.gov/~/media/ascr/pdf/programdocuments/docs/ascr-eod-workshop-2015-report_160524.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

Management, Analysis, and Visualization of EOD

http://science.energy.gov/~/media/ascr/pdf/programdocuments/docs/ascr-eod-workshop-2015-report_160524.pdf

Selected Findings and Excerpts:

- All Experimental and Observational Science (EOS) projects represented at this workshop struggle to keep up with the demands and opportunities that a flood of data offers.
- Scientific data is increasingly at risk of being unusable.
- EOS projects are impeded due to significant “data lifecycle” needs that are largely unmet.
- “. . . our only archival process right now is that provided by the published journal. ” (Neutron source case study)

Perspectives on Data (from Washington)

Policy and Implementation

https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

http://www.energy.gov/downloads/doe-public-access-plan

http://www.energy.gov/datamanagement/doe-policy-digital-research-data-management

http://science.energy.gov/funding-opportunities/digital-data-management/

http://www.osti.gov/pages/

Brief History

- **COMPETES 2010 “Interagency Public Access Committee”**
- Office of Science Working Group on Digital Data
- Office of Science FACA Reports (2011)
- OSTP Request for Information (2012)
- Office of Science User Facility Input (2013)
- **OSTP Memo “Increasing Access to the Results of Federally Funded Scientific Research” (Feb., 2013)**
- **DOE Public Access Plan and Office of Science Statement on Digital Data Management (July, 2014)**
- **DOE Policy for Digital Research Data Management (Sept, 2015)**

Office of Science Statement on Digital Data Management

http://science.energy.gov/funding-opportunities/digital-data-management/

U.S. DEPARTMENT OF ENERGY | Office of Science

Search SC Website | SC Site Search | GO

Funding Opportunities

Statement on Digital Data Management

Print | Text Size: A A A Feedback [+] | Share Page

The Office of Science mission is to deliver the scientific discoveries and major scientific tools that transform our understanding of nature and advance the energy, economic, and national security of the United States. The Office of Science Statement on Digital Data Management has been developed with input from a variety of stakeholders in this mission[1].

Here, data management involves all stages of the digital data life cycle including capture, analysis, sharing, and preservation. The focus of this statement is sharing and preservation of digital research data.

Table of Contents

Office of Science Statement on Digital Data Management

Principles

- Effective data management has the potential to increase the pace of scientific discovery and promote more efficient and effective use of government funding and resources. Data management planning should be an integral part of research planning.

- Sharing and preserving data are central to protecting the integrity of science by facilitating validation of results and to advancing science by broadening the value of research data to disciplines other than the originating one and to society at large. To the greatest extent and with the fewest constraints possible, and consistent with the requirements and other principles of this Statement, data sharing should make digital research data available to and useful for the scientific community, industry, and the public.

- Not all data need to be shared or preserved. The costs and benefits of doing so should be considered in data management planning.

Office of Science Statement on Digital Data Management

- Requirements apply to proposals for research funding
- Requirements apply to proposals submitted for new, renewal, and some supplemental research funding
- Requirements apply to proposals regardless of the PI's institution
- Requirements do *not* apply to applications to use Office of Science user facilities.
- SC sponsored research activities at the DOE National Laboratories for which a DOE-approved DMP does not already exist will be required to develop a DMP.
- DMPs will be reviewed as part of the overall Office of Science research proposal merit review process.
- Additional requirements and review criteria for the DMP may be identified by the sponsoring program or sub-program, or in the solicitation.

Office of Science Statement on Digital Data Management

http://science.energy.gov/funding-opportunities/digital-data-management/

Requirements (abridged)

1. DMPs should describe whether and how data generated in the course of the proposed research will be shared and preserved. DMPs must describe how data sharing and preservation will enable validation of results, or how results could be validated if data are not shared or preserved.

2. DMPs should provide a plan for making all research data displayed in publications resulting from the proposed research open, machine-readable, and digitally accessible to the public at the time of publication. The published article should indicate how these data can be accessed.

3. DMPs should consult and reference available information about data management resources to be used in the course of the proposed research. In particular, DMPs that explicitly or implicitly commit data management resources at a facility beyond what is conventionally made available to approved users should be accompanied by written approval from that facility.

4. DMPs must protect confidentiality, personal privacy, Personally Identifiable Information, and U.S. national, homeland, and economic security; recognize proprietary interests, business confidential information, and intellectual property rights; avoid significant negative impact on innovation, and U.S. competitiveness; and otherwise be consistent with all applicable laws, regulations, and DOE orders and policies. There is no requirement to share proprietary data.

Definitions

Digital Research Data:

The term *digital data* encompasses a wide variety of information stored in digital form including: experimental, observational, and simulation data; codes, software and algorithms; text; numeric information; images; video; audio; and associated metadata. It also encompasses information in a variety of different forms including raw, processed, and analyzed data, published and archived data.

This statement focuses on *digital research data,* which are *research data* that can be stored digitally and accessed electronically. OMB Circular A110 defines *research data* as follows:

> "Research data is defined as the recorded factual material commonly accepted in the scientific community as necessary to validate research findings, but not any of the following: preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues. This 'recorded' material excludes physical objects (e.g., laboratory samples). Research data also do not include:
>
> (A) Trade secrets, commercial information, materials necessary to be held confidential by a researcher until they are published, or similar information which is protected under law; and
>
> (B) Personnel and medical information and similar information the disclosure of which would constitute a clearly unwarranted invasion of personal privacy, such as information that could be used to identify a particular person in a research study."

Definitions

Data Preservation:

Data preservation means providing for the usability of data beyond the lifetime of the research activity that generated them.

Data Sharing:

Data sharing means making data available to people other than those who have generated them. Examples of data sharing range from bilateral communications with colleagues, to providing free, unrestricted access to the public through, for example, a web-based platform.

Validate:

In the context of this statement, *validate* means to support, corroborate, verify, or otherwise determine the legitimacy of the research findings. Validation of research findings could be accomplished by reproducing the original experiment or analyses; comparing and contrasting the results against those of a new experiment or analyses; or by some other means.

Additional Guidance and Requirements from Program Offices

Funding Opportunities

Grants & Contracts Support

Award Search

Find Funding

Early Career Research Program

Statement on Digital Data Management

Suggested Elements for a Data Management Plan

Frequently Asked Questions

Resources at the Office of Science User Facilities

Acknowledgements of Federal Support

CONTACT INFORMATION
Office of Science
U.S. Department of Energy
1000 Independence Ave., SW
Washington, DC 20585

Statement on Digital Data Management

Data Management Resources at the Office of Science User Facilities

Print | Text Size: A A A Feedback [+] | Share Page

Follow the links to learn more about data management resources at each of the Office of Science User Facilities. If you do not see your facility of interest listed here, please consult the appropriate Office of Science Program page.

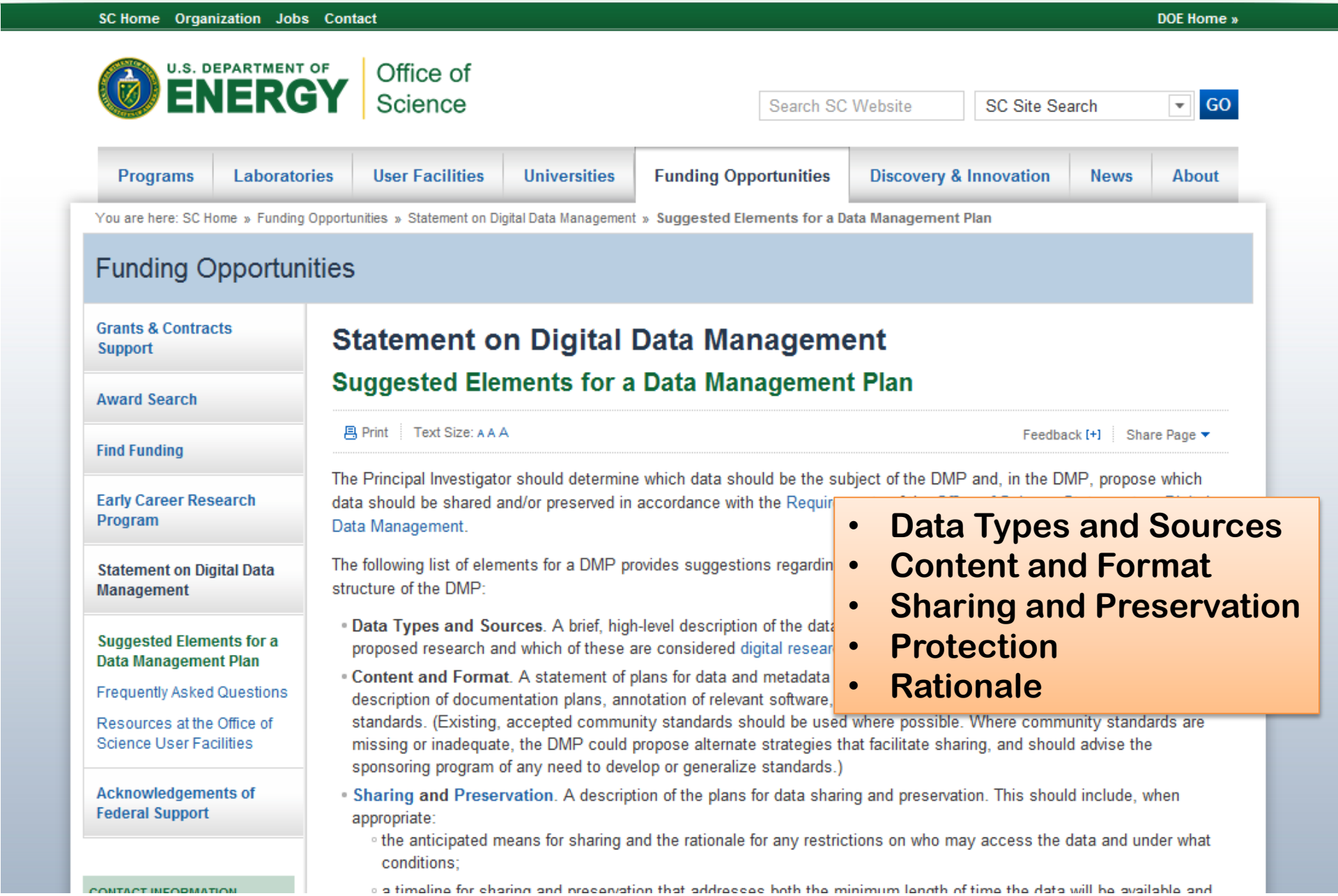
Advanced Scientific Computing Research (ASCR)

Facility	Host Institution	Data Management Resources
National Energy Research Scientific Computing Center (NERSC)	LBNL	Link
Argonne Leadership Computing Facility (ALCF)	ANL	Link
Oak Ridge Leadership Computing Facility (OLCF)	ORNL	Link
Energy Sciences Network (ESnet)	LBNL	Link

Basic Energy Sciences (BES)

Facility	Host Institution	Data Management Resources
Light Sources		
Advanced Light Source (ALS)	LBNL	Link
Advanced Photon Source (APS)	ANL	Link
Linac Coherent Light Source (LCLS)	SLAC	Link
National Synchrotron Light Source (NSLS)	BNL	Link

Suggested Elements for a Data Management Plan



- **Data Types and Sources**
- **Content and Format**
- **Sharing and Preservation**
- **Protection**
- **Rationale**

Data ID Service

http://www.osti.gov/home/doe-data-id-service

osti.gov
Speeding access to science information from DOE and beyond

HOME | ABOUT OSTI | SCIENCE SEARCH TOOLS | DOE PAGES Beta / PUBLIC ACCESS | COMMUNICATIONS | DOE STI PROGRAM

DOE Data ID Service

DataCite | Contact DOE Data ID Service

DOE Data ID Service

The DOE Office of Scientific and Technical Information (OSTI) offers a service for registering datasets to help increase access to digital data from DOE-funded scientific research. Through the DOE Data ID Service, OSTI assigns persistent identifiers, known as Digital Object Identifiers (DOIs), to datasets submitted by DOE and its contractor and grantee researchers and registers the DOIs with DataCite to aid in citation, discovery, retrieval, and reuse.

DOIs can be used to search for datasets on a number of OSTI's scientific and technical information database products. OSTI assigns and registers DOIs for datasets for DOE researchers as a free service to enhance the Department of Energy's management of this important resource.

OSTI, the only U.S. government member of DataCite, can assign DOIs to other federal agencies' datasets on a cost-reimbursable basis. For information about this interagency service, please contact DOE Data ID Service.

DOE Data ID Service Fact Card (654.74KB pdf)

SC Funding Opportunity Announcement Language

APPENDIX 6: DATA MANAGEMENT PLAN

Provide a Data Management Plan (DMP) that addresses the following requirements:

1. DMPs should describe whether and how data generated in the course of the proposed research will be shared and preserved. If the plan is not to share and/or preserve certain data, then the plan must explain the basis of the decision (for example, cost/benefit considerations, other parameters of feasibility, scientific appropriateness, or limitations discussed in #4). At a minimum, DMPs must describe how data sharing and preservation will enable validation of results, or how results could be validated if data are not shared or preserved.
2. DMPs should provide a plan for making all research data displayed in publications resulting from the proposed research open, machine-readable, and digitally accessible to the public at the time of publication. This includes data that are displayed in charts, figures, images, etc. In addition, the underlying digital research data used to generate the displayed data should be made as accessible as possible to the public in accordance with the principles stated in the Office of Science Statement on Digital Data Management (http://science.energy.gov/funding-opportunities/digital-data-management/). This requirement could be met by including the data as supplementary information to the published article, or through other means. The published article should indicate how these data can be accessed.
3. DMPs should consult and reference available information about data management resources to be used in the course of the proposed research. In particular, DMPs that explicitly or implicitly commit data management resources at a facility beyond what is conventionally made available to approved users should be accompanied by written

SC Funding Opportunity Announcement Language

Merit Review Criteria:

SCIENTIFIC AND/OR TECHNICAL MERIT OF THE PROPOSED RESEARCH

- What is the scientific innovation of proposed effort?
- How does the proposed work compare with other efforts in its field, both in terms of scientific and/or technical merit and originality?
- How might the results of the proposed work impact the direction, progress, and thinking in relevant scientific fields of research?
- What is the likelihood of achieving influential results?
- Is the Data Management Plan suitable for the proposed research and to what extent does it support the validation of research results?