Management, Analysis, and Visualization of Experimental and Observational Data – The Convergence of Data and Computing

E. Wes Bethel[1], Martin Greenwald[2], Kersten Kleese van Dam[3], Manish Parashar[4], Stefan M. Wild[5], and H. Steven Wiley[6]

[1] Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[2] Massachusetts Institute of Technology, Cambridge, MA, USA
[3] Brookhaven National Laboratory, Upton, NY, USA
[4] Rutgers University, Piscataway, NJ, USA
[5] Argonne National Laboratory, Lemont, IL, USA
[6] Pacific Northwest National Laboratory, Richland, WA, USA

August, 2016

Acknowledgment

This work was supported by the Director, Office of Science, Office of Advanced Scientific Computing Research, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231, program manager Dr. Lucy Nowell.

Legal Disclaimer

Management, Analysis, and Visualization of Experimental and Observational Data – The Convergence of Data and Computing

E. Wes Bethel*
Lawrence Berkeley National Laboratory

Martin Greenwald†
Massachusetts Institute of Technology

Kerstin Kleese van Dam‡
Brookhaven National Laboratory

Manish Parashar§
Rutgers University

Stefan M. Wild¶
Argonne National Laboratory

H. Steven Wiley‖
Pacific Northwest National Laboratory

Abstract

Scientific user facilities—particle accelerators, telescopes, colliders, supercomputers, light sources, sequencing facilities, and more—operated by the U.S. Department of Energy (DOE) Office of Science (SC) generate ever increasing volumes of data at unprecedented rates from experiments, observations, and simulations. At the same time there is a growing community of experimentalists that require real-time data analysis feedback, to enable them to steer their complex experimental instruments to optimized scientific outcomes and new discoveries. Recent efforts in DOE-SC have focused on articulating the data-centric challenges and opportunities facing these science communities. Key challenges include difficulties coping with data size, rate, and complexity in the context of both real-time and post-experiment data analysis and interpretation. Solutions will require algorithmic and mathematical advances, as well as hardware and software infrastructures that adequately support data-intensive scientific workloads. This paper presents the summary findings of a workshop held by DOE-SC in September 2015, convened to identify the major challenges and the research that is needed to meet those challenges.

1 Introduction

The Department of Energy (DOE) Office of Science (SC) operates dozens of national science user facilities that span many disciplines [9]. These facilities include accelerators, colliders, supercomputers, light sources, and neutron sources, as well as facilities for studying the nanoworld, genomes, the environment, the atmosphere, and the cosmos. In Fiscal Year 2014, over 33,000 researchers from academia, industry, and government laboratories, from all fifty states and the District of Columbia, utilized these unique facilities to perform new scientific research. Each of these facilities generates vast amounts of scientific data, and thanks to advances in technology, the size, rate, and complexity of this data is rapidly increasing. A growing concern is that advances in the science programs will not be able to keep pace with increasing data rates due to a lack of resources, the need for research and development of tools as well as with platforms and infrastructures needed to manage, analyze, and act on the growing collections of data. All of the above are needed to derive novel scientific insights from data.

With the goal of understanding, inventorying, and articulating the data-centric needs and challenges of the experimental and observational science (EOS) community in DOE-SC, the Office of Advanced Scientific Computing Research (ASCR) held a workshop, from 29 September 2015 through 1 October 2015 in Bethesda, MD. The workshop report [4] consists of inputs from representatives from a sampling of DOE-SC EOS facilities and researchers in mathematics and computer science. The findings of the workshop, drawn from detailed discussions of participants who reviewed a wide range of exemplary *science use cases*, indicate that there are acute and urgent needs regarding the management, analysis, and visualization of experimental and observational data (EOD) collected and generated by EOS at DOE-SC user facilities.

The science needs articulated in the workshop report, along with the findings, recommendations, and detailed discussion of issues, collectively are consistent with the vision articulated in the National Strategic Computing Initiative (NCSI) [19, 13] and the Big Data Research and Development Initiative [24, 18]. The workshop report also describes multiple opportunities for cultivating a research, development, and deployment path that will help realize that vision. Specifically, the science use cases reveal a trend toward the *convergence of data and computing*: data- and compute-centric needs and suggests that opportunities in these research areas are increasingly intertwined, interrelated, and symbiotic. Advances in our ability to collect data will require advances in computational capabilities to understand, preserve, share, and make optimal use of data, and can positively impact the quality and value of our science by improving the quality and reusability of the data we collect.

This paper makes two primary contributions to the eScience community. First, it contains a canonical science use case that captures many of the high-level characteristics of how large-scale EOS projects in DOE-SC make use of the EOD they collect. This canonical use case is a composite sketch drawn from eleven specific use cases provided by the DOE-SC EOS community. Second, this paper presents a set of data-centric challenges, distilled from the eleven science use cases in [4], that DOE-SC EOS projects face now and in the future. These two contributions are a summarization of ideas presented in the September 2015 ASCR EOD workshop report [4], which contains additional qualitative and quantitative information about the specific EOS projects.

We begin with a discussion of a canonical use case scenario (§2) to provide context and orientation to DOE-SC science facilities. That section lays the groundwork for the subsequent sections (§3–§9), which in turn present the primary findings of the workshop. We discuss related work (§10) before making concluding remarks.

*ewbethel@lbl.gov
†g@psfc.mit.edu
‡kleese@bnl.gov
§parashar@rutgers.edu
¶wild@anl.gov
‖steven.wiley@pnnl.gov

Table 1: The eleven use cases provided by the DOE-SC EOS community.

1	Environmental Molecular Sciences Laboratory
2	Climate Simulation and Analysis
3	Atmospheric Radiation Measurement Climate Research Facility
4	Advanced Light Source
5	Linac Coherent Light Source
6	Oak Ridge National Laboratory Neutron Sources
7	Scanning Probe and Electron Microscopies
8	Advanced Photon Source
9	Deep Underground Neutrino Experiment
10	Open Numerical Laboratories
11	DOE HEP Cosmic Frontier

2 A Canonical Science Use Case Scenario

The canonical science use case in this section is a composite of eleven use cases (see Table 1) submitted by the DOE-SC EOS community. Those eleven use cases span a diversity of science topics from the DOE-SC offices of Biological and Environmental Research (BER), Basic Energy Sciences (BES), and High Energy Physics (HEP). Prior to the workshop, EOS researchers from those areas provided detailed information specific to their EOS project(s) that answered the following set of questions:

Present- or near-term issues. Each EOS project provided a description of the science facility, how the facility or experiment "does science" with the EOD they collect, and a "flowchart" (verbal or pictorial) describing the data lifecycle starting with data acquisition and including all processing stages, all the way through dissemination.

Future issues. Each EOS project provided information from the same categories for both the present- or near-term issues as well as looking ahead, usually 3–5 years, into the future.

Data lifecycle. Each EOS project provided information about how data is used and the key issues that need to be addressed, at each of the primary data lifecycle stages, in both present and future scenarios.

Data requirements. For each stage in the data lifecycle, each EOS project provided information about data "speeds and feeds," throughput requirements, and specific data-centric capabilities needed for the specific science use case.

Impediments, gaps, needs, and challenges. Each EOS project provided a list of data-centric impediments or barriers they presently face and expect to face in the future.

The canonical use case below identifies the major stages in the data lifecycle and the types of processing that might take place there. The challenges faced by EOS projects in achieving these objectives are the subject of later sections.

Figure 1 is an illustration that gives a general overview of the major data lifecycle and processing stages in DOE-SC EOS projects. Although Figure 1 is a generalization of the eleven use cases in the workshop report, it captures all their major thematic elements. The following subsections discuss these major thematic elements.

2.1 Data Products

First and foremost, the primary objective for these EOS products is to support scientific research that generates large amounts of data from experiments and observations. In some cases, this EOD will be turned into "a product" that is then given to a single principal investigator (PI). In other cases, data products are created and shared by an entire community. In all cases, as elaborated below, these EOS projects have clear requirements for carefully capturing *provenance* (i.e., information about the conditions under which data is collected), what types of processing it underwent, and so forth, and preserving this information in the form of *metadata*. There are numerous challenges associated with growing data size and complexity, and all EOS projects point to the fact that EOD may have a long lifespan, which in turn creates challenges regarding long-term data storage and dissemination. These challenges are all the subject of later sections in this paper.

2.2 Use of HPC Computing Facilities

The background in Figure 1 is shaded to indicate that EOS projects may do some data processing "close to" the instrument at the science user facility, while some of the processing requires, due to data size or other factors, access to and use of large-scale HPC computing facilities. The balance point, between computing at science user facilities and at HPC facilities, highly varies from project to project and depends on specific project needs. In later sections, we delve into the challenges of EOS projects' use of HPC computing facilities, which span diverse topics ranging from meeting time-sensitive computational requirements, to long-term data storage and dissemination.

2.3 Data Processing, Analysis, and Understanding

Hamming's statement that "the purpose of computing is insight, not numbers" [14] applies equally to EOS projects and their use of data. Figure 1 shows that operations on data, which include various types of processing, such as filtering, reduction, cataloging, analysis, and provenance capture, occur at multiple stages in the data lifecycle.

2.4 Data Sharing and Collaboration

Although it may not be readily apparent from Figure 1, the workshop's science use cases reveal a theme that collaboration and sharing are central to EOS science. Data produced by EOS projects is almost always turned into data products that are disseminated to a single PI, to a small group, or to a much larger, geographically distributed community. To be useful, however, such data needs to be curated and made easily understandable and accessible by all users.

Efficient use of shared data requires adequate metadata, if only to inform the community about the precise conditions under which it was collected. Typically, any given data set will be interpreted with respect to prior knowledge or will be combined with previously collected data. Thus, EOS data sets are usually a part of a much larger research study in which the metadata needed to set the context of the data is as important as the data itself. In addition, almost all EOS project data must be extensively processed by specialized software packages to convert the raw instrument data into a form that can be used in broader scientific studies.

2.5 Uses of EOD

Although Figure 1 shows a generalization of major data lifecycle stages, the paragraphs below summarize a few of

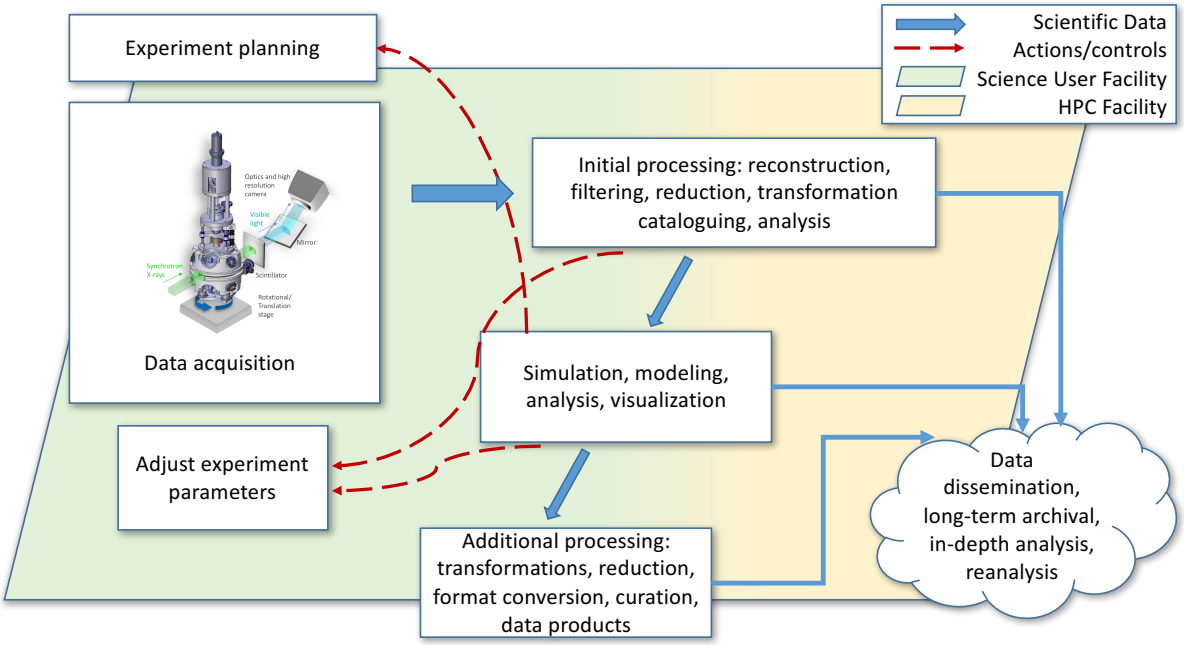

Figure 1: An illustration showing the various stages in the EOS data lifecycle.

the many uses of EOD identified in the eleven scientific use cases presented at the workshop.

Data products from EOS. In the early stages of experiment planning, a single PI (or community of investigators) tends to focus on a particular science mission or objective. The experiment is conducted, and the "first light" output of these experiments consists of a data product of some type. Here, the phrase "data product" could mean something relatively narrow and focused, such as a lab analysis of a sample, or it could mean something much broader, such as a large collection of data from multimodal sources like a sky survey.

Primary analysis. With the data product(s) in hand, EO researchers then proceed to perform in-depth analysis as they focus on hypothesis testing, verification that the experiment went as intended, and/or any number of other tasks within the primary mission's focus.

Some of this analysis processing may require non-trivial steps, including assimilation with simulation or model-based approaches and statistical inverse problems, in order to extract key measurements from the EOD. These analysis results may produce outputs that could become data products in their own right, such as catalogs of features or properties extracted from the EOD.

Reanalysis. After an experiment and its initial study are completed, the resulting primary and supplementary data products may have a long lifespan. Over time, there may be the possibility for new scientific inquiry focusing on new hypotheses not envisioned at the time of the original experiment. For example, the Sloan Digital Sky survey (first light in 1998 [22]) is still operating today and its data are actively used by a worldwide science community.

Reference data sets. Once generated, a given data set may serve as an invaluable reference in many different ways in the future. Here, a significant issue is that not every EO data set will be equally used by the public, and over the years some of them will fade into irrelevancy while others emerge as a community reference. It is these latter data sets that need to be kept for a long time, even if just for comparison and reference. For such collections, which are used by many different refereed publications, reproducibility of these analyses will become another reason to keep the data, even when better and higher resolution alternatives become available.

Broad dissemination of results. In 2013, the Office of Science and Technology Policy issued a memorandum directing federally funded programs to make the results of their research freely available to the public, generally within one year of publication [12]. This broad, public access to research data of all types, and EOD in particular, could have far-reaching impacts. It could entail uses of EOD that go far beyond those envisioned in our 2015 workshop report.

3 Challenges of Exploding Data Size, Rate, Complexity, and Diversity

Data size and rate of collection at science user facilities are growing at a rapid rate. Each of the eleven EOS use cases provide details about expected and anticipated growth in data rates. One of the primary drivers for increasing data size is the increase in the resolution of the instrument sensors and detectors. From these summaries of projected growth rates, we see a near future where individual facilities, of which there are dozens, are each generating collections of data in the range of tens to hundreds of petabytes per year. These projections suggest, when integrating across the entire program, that these *science user facilities will be soon collectively acquiring exabytes of data per year.* Affordable data storage, effective data access, distribution, and curation, and meaningful analysis are key challenges that these facilities increasingly face.

All EOS projects represented at this workshop are having difficulty coping with with the demands and opportunities

that the flood of data offers. The complexity of the data, new challenges in analytics and visualization, difficulties in capturing sufficient metadata, and ease-of-use problems are impediments to use and adoption of many types of data-centric tools and infrastructure, hampering the effort to harness the wealth of data in the service of scientific discovery.

> A key limitation today is our [in]ability to analyze and visualize the acquired data due to its volume, velocity and variety.
> –*Advanced Photon Source [3]*

The problems associated with data size and velocity are compounded when a given EOS community relies on multiple instruments, such as the DOE-SC's Cosmic Frontier projects, which carry out sky surveys (e.g., the Sloan Digital Sky Survey [22]) using multiple instruments. These are expected to produce data estimated to be on the order of hundreds of petabytes. These data need to be available to the research community over a long period of time. Survey data with a long shelf life can be very valuable because they can be mined and analyzed in many different ways, thereby providing a stable resource for developing new approaches for data exploration and analysis. in addition to the intrinsic science value.

The growth in data volumes creates challenges beyond issues of processing and storage, but also for data transfer, particularly in experiments that rely on real-time feedback to the instrument operator. This problem is complicated even further by the fact that many experiments can be run concurrently with parallel data flows coming from independent detectors.

Another challenge associated with increased data size and complexity is the need to support data integration and data discovery through the appropriate collection and management of metadata and derived data products associated with experimental and simulation studies.

An ongoing concern in EOS projects is the need to detect errors in automated, high-throughput workflows. There is a clear convergence here between computing and data, where computational methods can be brought into the picture to ensure the best possible data are collected during an experiment. In some cases, errors occur during data acquisition. These errors can be mitigated and/or corrected after taking data through advanced algorithms that can model the dynamical effects of the acquisition instrument to produce a data set with minimized and/or quantified error.

A related concern is the loss of science and opportunities for science discovery due to data loss. One example is the Cosmic Frontier projects, where data loss can occur in studies of transients because of possible inefficiencies in detection technology, classification algorithms, and lack of follow-up resources. Other issues that prevent making use of the complete data set are technical issues such as lack of understanding of foregrounds, incomplete modeling of the atmosphere, and detector noise.

Although coping with the increasing size and rate of data inflows from experiments and observations is challenging, there is a corresponding set of issues at the other end of the data pipeline. EOS projects typically also produce data products that are derived from raw EOD and, in many cases, from the results of numerical calculations. Some data products are produced for individual users, while other data products are intended to be used by entire communities or as reference data sets. As mentioned above, it is necessary to have a clear record of information (metadata) about the data for these data products to be useful and usable, as well as a long-term plan for addressing the archiving, curation, and Dissemination of these data products.

4 Science Use of Large-Scale HPC Facilities

EOS projects' use of tools and facilities, which are designed for HPC workloads, have realized varying degrees of success. Meeting the challenges of the explosion of data from EOS projects requires computational platforms, networking, and storage of greater capacity and lower latency, along with software infrastructure suited to their needs. However, existing HPC platforms and software tools are designed and provisioned for high-concurrency HPC workloads, single-project data products, and comparatively simpler data needs. Future HPC architectures are expected to be more I/O-challenged.

EOS projects look to large-scale HPC computing facilities to help serve workloads that can be characterized as: requiring fast turnaround for computing tasks, having processing pipelines that are distributed in nature and involve the movement of very large amounts of data, long-term storage of data, as well as providing access to data to a potentially diverse set of stakeholders and consumers.

> Each beamline operates with unique capabilities and an independent scientific mission. . . . Computational needs and strategies may differ considerably across beamlines, but computation is required for nearly every aspect of the facility.
> –*Advanced Photon Source*

The issue of fast turnaround is so significant that we expand on related findings below. In brief, EOS projects using instruments such as beamlines, require computational resources within minutes, or perhaps seconds, of when data are generated, which is incompatible with the queued structure employed on today's leadership-class HPC machines.

The EOS science use cases presented at the workshop describe variants of data handling and processing activities that can be characterized as distributed computing models. The typical design pattern involves first collecting data at the instrument, performing some processing close to the instrument, then moving data to a large-scale facility for more lengthy calculations and preparation of data products, followed by dissemination of data products. The way each project implements this pattern varies according to their needs and available resources.

As another example, the Deep Underground Neutrino Experiment (DUNE, [8]) experiment presently uses a combination of local, on-site computing and HPC facilities at Fermi National Accelerator Laboratory, which also is expected to host a full replica of data recorded by the prototype instrument. DUNE plans to keep full data replicas off site for redundancy, as well as to opportunistically leverage computing resources, including those outside of DOE-SC. DUNE is targeting the design and development of project-wide software infrastructure designed to maintain portable and accessible software that can be used at any particular institution and run transparently on modern grid and/or cloud resources as part of a distributed-processing, data-centric workflow.

> Procedures for moving data from place to place, including tools for automating [a] resilient workflow for orchestrating distributed data-related operations are a bottleneck.
> –*Accelerated Climate Modeling for Energy project [1]*

There is a clear need for community- or facility-centric data repositories for data storage and dissemination with sufficient bandwidth and with an easy interface for interacting with tools for data analytics. The most useful platforms need to be massively parallel to support a combination of visualization and analysis tools. It is very likely that DOE facilities (both HPC facilities and science user facilities) will take on a significantly larger role in data archiving, transfer, and analysis. It is also possible that commercial cloud resources could become a major resource in these areas—although several outstanding questions remain (e.g., cost models, data archiving and transfer); this disruptive possibility needs to be continuously explored. The main new hardware trend of interest for DOE facilities—in the relatively near-term—is the evolution and integration of HPC systems within a data-centric usage model.

The needs of data sharing sites are quite distinct from those simply designed to store or analyze data. Data sharing and dissemination software must have robust features in searching for specific data types and for linking the data to people, studies, scientific fields, and published results.

5 Time-Sensitive Computing

Many projects have time-critical data needs (e.g., arising from human-in-the-loop experiment steering). These projects require a low-latency, high-throughput infrastructure for data movement, analysis, processing, and storage. However, computing platforms currently available to such researchers are insufficient in capacity or turnaround. The lack of large and capable facilities tuned to EOS needs are common across many disciplines.

Many EOS projects use, or hope to use, large-scale HPC platforms and high-speed networking to do real-time processing of experiment data. The desire is for high-throughput, fast turnaround to enable adjustment of experiment parameters while the experiment is in progress, thereby maximizing the scientific outcomes of the experiment.

For example, at the very first stages of the analysis workflow, scanning electron microscopy projects—such as those at the Center for Nanophase Materials Sciences [5]—are interested in collecting full detector response at the fastest meaningful rates in order to assess tool performance and adjust parameters on-the-fly. Fast visualization schemes would also be useful for monitoring samples and quality of output signals.

6 The Risk of Unusable Data

Scientific data is increasingly at risk of being unusable, untraceable, or unreproducible. Without adequate metadata and provenance, scientific data has limited usefulness because its origins are undocumented and unknown, thereby limiting the ability to validate results or to make use of such data for other purposes. However, today the capture of these critical information often relies on manual, non-digital and non-sharable approaches, hindering scientific discovery particularly in increasingly high-velocity, high-volume data environments.

In some projects, data-centric operations, which involve management, analysis, movement, and distribution, are the responsibility of an individual user, with whatever limited knowledge and capability is available to them. As a result, only a fraction of collected data is ever analyzed, and only a fraction of that data is ever published and made available for community-wide use.

One very real consequence of current data management systems is that most of the data is difficult to use by anyone other than the original group that generated it. This problem must be solved if making data publicly available is intended to have any useful purpose. In addition, much necessary metadata is never collected because of the lack of understanding of what is required for data sharing by the primary investigator(s) and the lack of easy-to-use tools to capture it. The overall cost and complexity of metadata recording and consolidation is currently prohibitive, which is a primary reason why metadata is rarely collected. Unfortunately, this means that the associated data cannot be easily exchanged or reused. Systematic collection of the metadata that describes the provenance of stored data is typically inadequate, limiting the integrity, traceability, and reproducibility of research products.

> . . . relevant data should be made available to the scientific community after some amount of time. But more than data preservation is required—proactive data curation is necessary for the data to be really useful. . . . The benefit of curation would be to reduce duplication of effort in data creation, but also for the re-use of data for further high quality research.
> –*Advanced Light Source [2]*

There is interest in having access to data after the current research is published. Such access needs to ensure that enough metadata is stored so that the data can be analyzed appropriately. Although certain classes of metadata (e.g., the who, the when) can be captured automatically, there is a need to capture the reason why certain aspects of an analysis or data transformation or reduction operation was performed. This information needs to be provided by the data generators and archived with the data so that subsequent access is useful, and can be utilized by researchers beyond the group that acquired it originally.

7 The Central Role of Collaboration and Sharing

Collaboration and sharing of data, tools, and methodologies are central to modern EOS projects, yet there is insufficient infrastructure to facilitate such interactions. Common tools and methodologies for sharing and collaboration in data-intensive sciences have not been widely developed, deployed, or adopted. The limit is generally not simply data transfer, but rather a lack of widely-used tools to produce and consume well-characterized data collections that include the desired level of annotation, metadata, and provenance. Collaborations also require an ability to share software tools, source code, data models, and formats, and to provide workflows that are reproducible. Beyond established collaborations, there is a clear need to share tools and approaches between groups and disciplines to minimize the unnecessary duplication of effort. In many cases, existing tools are inadequate or too difficult to use.

By their nature, the mission focus for EOS projects is to collect data, and to share it. This theme is present in all the use cases presented at the workshop. The projects differ in some key ways: some projects' immediate focus is on sharing data with a primary principal investigator or group, while others focus on sharing data with larger communities. Although making data accessible for download over the Internet lowers the barrier to accessibility for a potentially large number of consumers, doing so is only a small part of a larger landscape of collaboration and sharing.

> Understanding the process of how science is actually done, what information needs to be captured and where the data is generated are key issues that must be addressed to enable effective data sharing.
> –*Environmental Molecular Sciences Laboratory [11]*

The use cases provide several compelling reasons why collaboration and sharing is important. First, sharing software has the potential effect of reducing costs, particularly of software development. The idea is that redundancy of effort—software development—is reduced when key methods and tools can be reused across different projects. Sharing data, particularly curated data, would be to reduce the duplication of effort in data creation, as well as for data re-use for further high-quality research. Another benefit would be that it could lead to more algorithms and software being made available to the community, as researchers write code that can be benchmarked and used against curated data.

One concept that is central to achieving the ability to share data and tools is the idea of community-centric, or "standard" data models and formats for both data and metadata. The climate community, for example, has realized a degree of success in sharing data as well as software tools for working with data, due to its use of a data model or format that has broad community support. This idea is identified as a need or an impediment in several of the science use cases.

> Current technologies are inadequate for sharing [. . .] data between group members. The EOS community needs a more fluid means for sharing data and working together.
> –*Environmental Molecular Sciences Laboratory*

The use cases identify several different ideas that are needs for or impediments to collaboration and sharing. One is that the issue of data and software sharing does not have program-wide visibility. As a result, progress in this space is *ad hoc*, with solutions for distributing data or software emerging on a per-facility or per-PI basis, with little or no coordination. Thus, there are many different sources of data and software, resulting in duplication of effort as well as a high barrier to finding data or software. Impediments that are most detrimental are related to the issues of data sharing and collaborating in large groups, methodological transparency, and dissemination and archival capabilities. Data and/or software that is "custom" and not curated is unlikely to be widely used. Better methods—interfaces and software tools, infrastructure—are needed to search and subset data without having to download an entire data set.

> Impediments that are most detrimental are related to the issues of data sharing and collaborating in large groups, methodological transparency, and dissemination and archival capabilities.
> –*Environmental Molecular Sciences Laboratory*

There is a deep interplay between the topic of collaboration, and the related but orthogonal topics of the overall data lifecycle, the usability of data and the associated challenges of metadata/provenance capture and long-term data archival and curation, and EOS's use of computing and data facilities. The interactions between these different focus areas is made more challenging by the rapid rate of growth in data size and the rate of data acquisition. Stated differently, successes in these related areas are building blocks for success in all areas of collaboration and data sharing.

8 Data Lifecycle

The term data lifecycle refers to all stages of data collection, movement, processing, analysis, management, curation, and sharing. Data collected by observation or experiment has a potentially long lifespan, and a potentially large set of consumers, but there presently is no solution or approach for data curation, quality management, and long-term distribution within DOE-SC that is generally and broadly applicable. At the same time, data retention policies at DOE-SC computing facilities are not designed for long-term retention nor for widespread dissemination.

EOS projects have "data lifecycle" needs that are significant, well defined, and that go well beyond what is provided by the current set of programs and projects in the ASCR computing facilities and research portfolio. EOD can have a long lifespan, yet there is no program-wide view or approach for the long-term curation, storage, and dissemination of such data; one EOS project indicated that it relies on whatever capabilities are provided by journals in association with publications as its solution to this problem.

> . . . our only archival process right now is that provided by the published journal.
> –*Spallation Neutron Source [23]*

Two key motivations for retaining data sets for a long period of time are for having a reference data set for use in evaluating the effectiveness of new methods over time, and for the opportunity for new discoveries not originally foreseen at the time the data was collected. Over the years, some data sets produced by simulation will emerge as a community reference. For such collections, which are likely to be used by many different authors in refereed publications, reproducibility of these analyses will become another reason to keep the data, even when better and higher resolution alternatives become available. For example, in tomography, the resulting tomogram is of comparable size to the raw images, which are also usually retained for the purposes of comparing the results of different tomographic reconstruction algorithms.

Results from projects like sky surveys may initially be focused on a few key science missions, but over time, a diverse set of science activities can be carried out with substantial discovery potential.

Current strategies for managing (accessing, processing, and keeping track of) the large number of data products are awkward at best, requiring a combination of methods. Science user facilities like the Advanced Photon Source (APS, [3]) do not provide a centralized and robust long-term data archive, since this service is categorized as a user responsibility. Most science user facilities have no explicit method for long-term archival and curation, and this is identified as an impediment. It is likely that in the future, science user facilities will be called upon to provide long-term storage and archival services. One stop-gap approach for long-term archival is to rely on the infrastructure and policies provided by the journal where a given paper is published. A welcome addition in the data universe would be a centralized DOE-SC facility that provides a mechanism for data archiving and retrieval, that could be provided as an option to users at low or no cost.

> Providing more access to the data, in a manner that can be used by more scientists, will improve efficiency, increase the impact of the science, and result in more papers per experiment.
> –*Spallation Neutron Source*

The issues related to data lifecycle management are broad, and cut across many different areas. We have identified challenges and research needed in areas germane to this topic: the automation of processing stages and automated data movement in EOS, data storage and retrieval, metadata and provenance, software engineering and infrastructure, data curation, collaboration, and interaction with computing service facilities.

9 The Central Role of Software

Software is a critical element for all EOS projects in all aspects of working with data and in meeting the challenges of increasing data size and complexity. Software is used for collecting, processing, and analyzing data, for preparing data products, and for automating complex multi-stage operations that may span distributed resources.

An important outcome of the workshop is the recognition of common needs across all the science domains. Although the computing needs of EOS projects vary from one project to the next, all EOS projects need computing and data storage/dissemination, along with a sustainable software ecosystem that can evolve over time to accommodate its data-centric requirements. This finding suggests that priority attention should be directed toward approaches that develop and support solutions that can be widely used by many EOS projects and facilities.

Software methods need to be capable of supporting the time-critical analysis and display tasks summarized above. Software methods, such as advanced algorithms for analysis, play a key role in improving the quality of data collected during an experiment, thereby improving the efficiency and quality of science. The idea is that EOS projects like beamlines require computational resources within minutes or perhaps seconds when data are available and cannot abide with the queued structure employed on leadership machines.

Because of the central role that software plays in nearly all aspects of working with data, EOS projects are particularly vulnerable to inefficiencies and increased costs resulting from inadequate software. For example, time inefficiencies result when data-centric pipelines and data movement activities must be executed manually rather than being automated and resilient. Inefficiencies in cost can result when a customized software component is created for one user but is not readily customizable or applicable to other users in the same facility, or across other science facilities.

> The biggest challenge to the facility is how to create the scientific software needed to run it: software for improving the experimental process; for implementing beamline data movement and reduction workflows; to perform preliminary quality assurance, visualization and reduction; for data analysis and interpretation; for automating analysis workflows and distribution to users.
>
> The most serious impediment the APS encounters is a lack of a DOE-wide view of software needs across the BES mission. Since each lab has its own portfolio of responsibilities, it devotes resources to those goals.
> –*Advanced Photon Source*

Software technology also plays a key role in encapsulating complexity and as an enabling technology. EOS projects want and need to be able to make use of advances in computational architectures, such as using HPC platforms for performing data-centric operations on larger data and with a faster turnaround. However, developing software for those platforms is often beyond the reach of a typical scientist-developer who may not have HPC software development skills. When it comes to the development of HPC code, there are fewer tools that ease the process for scientist-software developers (as opposed to computational experts) to transition from prototype code to HPC production code. The same idea extends to other areas of technology, such as creating data-centric pipelines that span distributed resources.

Increasingly, both simulations and experimental data analysis are elements of integrated workflows, which should resiliently automate key components of the data-handling pipeline, from collection to processing, analysis, archival, and dissemination. Many contemporary EOS projects articulate the need to combine computing with the experiment in real time, so as adjust experimental parameters on-the-fly to obtain the best possible data and science result from the experiment. Software methods need to be capable of supporting the time-critical analysis and display tasks summarized above. Meeting these challenges will require more powerful computing and networking infrastructure combined with a capable, robust and sustainable software ecosystem focusing on EOS needs. The flood of data available now and in the near future presents an opportunity that can be met only through concerted, coordinated, and sustained efforts to improve the software tools, methods, and facilities (computing, data) available to the EOS community.

Software is "digital data" that needs to undergo the rigors of curation, in the same way as do data from experimental and observational sciences, to facilitate its long-term archival

preservation and widespread dissemination. To be useful, software, like other forms of digital data, needs to have associated metadata along with documentation and examples of use. To be long-lived, it needs to be supported, maintained, and disseminated, something that is often not a part of the cost model. Wide-scale adoption of common tools will only occur if users are convinced that those tools will be maintained. Several use cases pointed to the desire to distribute software together with data, to expand the usability of data and to promote the repeatability of results as well as to promote the use of reference data and methods. One use case pointed out that their only practical avenue for doing so was to rely on the archival capabilities provided by the journal where results are published. The issues and motivations related to software curation, preservation, and dissemination are similar to those for other types of scientific data.

10 Related Work

The increasingly important role of data and data understanding in science is well recognized [16] along with attendant challenges [17]. Within the scientific community, there have been concerted efforts to capture data-centric challenges and requirements. In 2004, the *The Office of Science Data-Management Challenge* report [10] suggests that successfully addressing challenges related to the management and understanding of data one of the primary obstacles facing modern science.

These topics were revisited a decade later in 2013 in the *Data Crosscutting Requirements Review* [15], with many of the same observations as the 2004 report, but that also draw distinctions to the differences and similarities between data-centric issues specific to the sciences and those elsewhere, such as finance, health care, and so forth. Also in 2013, a National Research Council report *Frontiers in Massive Data Analysis* [6] summarized analysis challenges in particular, and highlighted the cross-disciplinary knowledge required to address these challenges. These issues are not unique to Department of Energy science areas. The biological and medical sciences, for example, have also grappled with issues of large-scale distributed data sets that require significant computational resources for their analysis along with potential approaches [20] that are similar to those described in the 2015 EOD report.

From 2015, *The Future of Scientific Workflows* report [7] examines data issues from a distributed computing perspective, where scientific data must be handled and acted on as part of an end-to-end workflow, which may involve processing stages cited at multiple geographic locations. The 2015 EOD report, which is the subject of this paper, takes a broader view than these earlier reports, to include the symbiotic relationship between the science user facilities and computational and networking infrastructure, along with key research targets in mathematics and computer science, that would come into play to service those data-centric workloads and meet scientific knowledge discovery challenges.

The increasing interplay between high-performance computing and large-scale data challenges is the subject of a recent article by Reed and Dongarra [21]. They suggest that advances in scientific research require growth in infrastructure for both computing and for analyzing data. They also point out the divergence in software architectures used for traditional HPC computing and for Big Data handling, due in part to the economics of widespread use of commodity components in industry.

The 2015 EOD report provides perspective on some of the challenges facing data-intensive science use of HPC facilities that have historically serviced computationally focused workloads. These challenges go far beyond what processors, interconnects and software components are in the system, and includes issues related to how knowledge is gained from data, how data is specifically used by several different data-intensive EOD science projects, and the specific impediments standing in the way of scientific progress in the face of a daunting deluge of scientific data.

11 Conclusions and Summary

Like many other areas of science, the science user facilities and projects operated by the U.S. Department of Energy's Office of Science are challenged by an increasing deluge of scientific data. The focus of this paper is on summarizing some of the key findings of a workshop convened in September 2015 to identify key challenges and new research and development needed to meet those challenges. Whereas this paper focuses on the key findings and challenges, that workshop report provides considerably further depth to the findings, as well as information contributed by workshop participants that articulates new research and development needed to meet those challenges.

The 2015 EOD report contains a diversity of specific recommendations for next steps towards meeting data-centric needs. Space constraints here limit a substantial discussion of these recommendations. In brief, they encompass both high- and detailed-level views of the problem space. For example, high-level issues focus on program-wide coordination of responses and efforts so as to overcome the fragmented and duplicative nature of how science user facilities approach cultivating solutions to data-centric challenges. Detail-level views focus on, for example, the need for new computational and mathematical methods for approximate and fast modeling and analysis calculations to meet the emerging use case of time-constrained use of HPC facilities so as to tune experiments on-the-fly with the objective of obtaining higher-quality experimental data.

Realizing success in meeting many of the data-centric challenges discussed here will likely entail efforts that require program-wide coordination and visibility. Many of the problems are simply too large and too far reaching to be tackled by an individual investigator or individual science user facility. For example, the objective of long-term data archival and dissemination is something that is needed by all EOS projects, and is something that requires significant forethought and attention to how EOD may be used in the future. This observation is applicable far beyond the set of science use cases represented in the workshop's report, and can help guide and shape how data-centric problems in all sciences might be approached elsewhere.

The science use cases in our workshop report reveal a trend toward the *convergence of data and computing*. The uses cases articulate both data- and compute-centric needs, and suggest that opportunities in these research areas are increasingly intertwined, interrelated, and symbiotic. Advances in our ability to collect data will require advances in computational capabilities to understand, preserve, share, and make optimal use of data, and can positively impact the quality and value of our science by improving the quality and reusability of the data we collect.

Acknowledgments

We are grateful to more than 50 workshop participants and 16 other co-authors for their invaluable contributions to the report summarized in this paper. This work, and the associated workshop, was supported in part by the Director, Office of Science, Office of Advanced Scientific Computing Research, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231, program manager Dr. Lucy Nowell.

References

[1] Accelerated Climate Modeling for Energy (ACME). http://climatemodeling.science.energy.gov/projects/accelerated-climate-modeling-energy/, last accessed May 2016.

[2] Advanced Light Source. https://www-als.lbl.gov/, last accessed May 2016.

[3] Advanced Photon Source. http://www.aps.anl.gov/, last accessed May 2016.

[4] E. Wes Bethel and Martin Greenwald (eds.). Report of the DOE Workshop on Management, Analysis, and Visualization of Experimental and Observational Data – The Convergence of Data and Computing. Technical report, Lawrence Berkeley National Laboratory, Berkeley, CA, USA, 94720, May 2016. LBNL-1005155.

[5] Center for Nanophase Materials Sciences. https://www.ornl.gov/facility/cnms/, last accessed May 2016.

[6] National Research Council. *Frontiers in Massive Data Analysis*. The National Academies Press, Washington, DC, 2013.

[7] Ewa Deelman and Tom Peterka (editors). The Future of Scientific Workflows, April 2015. http://science.energy.gov/~/media/ascr/pdf/programdocuments/docs/workflows_final_report.pdf.

[8] Deep Underground Neutrino Experiment. http://www.dunescience.org/, last accessed May 2016.

[9] US DOE. Office of Science User Facilities. http://science.energy.gov/user-facilities/user-facilities-at-a-glance/, last accessed May 2016.

[10] Richard Mount (editor). The Office of Science Data Management Challenge – Report from the DOE Office of Science Data Management Workshop Series. Technical report, Stanford Linear Accelerator Center, 2004. SLAC-R-782, http://www.er.doe.gov/ASCR/ProgramDocuments/Docs/Final-report-v26.pdf.

[11] Environmental Molecular Sciences Laboratory. https://www.emsl.pnl.gov/, last accessed May 2016.

[12] Expanding Public Access to the Results of Federally Funded Research. https://www.whitehouse.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research/, last accessed May 2016.

[13] FACT SHEET: National Strategic Computing Initiative, July 2015. https://www.whitehouse.gov/sites/default/files/microsites/ostp/nsci_fact_sheet.pdf, last accessed March 2016.

[14] Richard W Hamming. *Numerical Methods for Scientists and Engineers*. Dover Publications, Inc., New York, 2nd edition, 1986.

[15] Bruce Hendrickson and Arie Shoshani (editors). Data Crosscutting Requirements Review. Technical report, U. S. Department of Energy, Office of Science, April 2013.

[16] Tony Hey, Stewart Tansley, and Kristin Tolle, editors. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, Redmond, Washington, 2009.

[17] H. V. Jagadish, Johannes Gehrke, Alexandros Labrinidis, Yannis Papakonstantinou, Jignesh M. Patel, Raghu Ramakrishnan, and Cyrus Shahabi. Big data and its technical challenges. *Communications of the ACM*, 57(7):86–94, 2014.

[18] Tom Kalil and Fen Zhao. Unleashing the power of big data, April 2013. https://www.whitehouse.gov/blog/2013/04/18/unleashing-power-big-data/, last accessed March 2016.

[19] Barak Obama. Executive Order – Creating a National Strategic Computing Initiative, July 2015. https://www.whitehouse.gov/the-press-office/2015/07/29/executive-order-creating-national-strategic-computing-initiative/, last accessed March 2016.

[20] Russell A Poldrack and Krzysztof J Gorgolewski. Making big data open: data sharing in neuroimaging. *Nature Neuroscience*, 17(11):1510–1517, October 2014.

[21] Daniel A. Reed and Jack Dongarra. Exascale Computing and Big Data. *Commun. ACM*, 58(7):56–68, June 2015.

[22] Sloan Digital Sky Survey. http://www.sdss.org/, last accessed May 2016.

[23] Spallation Neutron Source. https://neutrons.ornl.gov/sns/, last accessed May 2016.

[24] Rick Weiss and Lisa-Joy Zgorski. Obama Administration Unveils "Big Data" Initiative: Announces $200 Million in New R&D Investments, March 2012. https://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf, last accessed March 2016.