



U.S. DEPARTMENT OF
ENERGY

Office of Science

Data Crosscutting Requirements Review

April 4-5, 2013

SPONSORED BY THE OFFICE OF ADVANCED SCIENTIFIC COMPUTING RESEARCH

Table of Contents

1 Executive Summary	5
2 Introduction	9
3 Scientific Drivers	11
3.1 High Energy Physics.....	11
3.1.1 The Energy Frontier	11
3.1.2 The Intensity Frontier.....	12
3.1.3 The Cosmic Frontier.....	14
3.2 Basic Energy Sciences.....	15
3.2.1 Data-intensive Computing for Light Sources.....	15
3.2.2 Emerging Data Challenges at the Nanoscience Centers	17
3.2.3 Data-intensive Neutron Facilities	19
3.3 Biological and Environmental Research.....	21
3.3.1 Climate Science Challenges and Motivation.....	21
3.3.2 Computational and Data Challenges for High-throughput Genome Science.....	23
3.3.3 The Environmental Molecular Sciences Laboratory.....	27
4 Crosscutting Computer Science and Mathematics Challenges	31
4.1 Data Processing	31
4.1.1 Data Acquisition	33
4.1.2 Data Reduction	33
4.1.3 Data Transformation	33
4.1.4 Data Movement	34
4.1.5 Workflows	34
4.1.6 Metadata and Provenance.....	35
4.2 Data Management	35
4.2.1 Data Movement (I/O)	36
4.2.2 Data Sharing	36
4.2.3 Data Retention and Curation	37
4.2.4 Search and Discovery	38
4.2.5 Storage.....	39
4.2.6 Data Models and Schema	39

4.3 Data Analysis	40
4.3.1 Data Quality	41
4.3.2 Improved Statistical, Machine Learning, and Image Analysis Algorithms.....	41
4.3.3 Approximate and Automated Algorithms	43
4.3.4 Multi-sensor and Multi-resolution Analysis	44
4.3.5 Visualization.....	44
4.3.6 Scalable Parallel Algorithms	46
4.3.7 Experimental Design	47
4.3.8 Inverse Problems	48
4.3.9 Inference, Prediction, and Reasoning under Uncertainty	49
5 Global Themes	51
5.1 Cost-model-based Data Processing System Design and Operation Optimization	51
5.2 Human Computer Interface.....	51
5.3 Software Quality, Resilience, and Readiness	51
5.4 Enhance Use of Modeling and Simulation for Experimental Design	52
5.5 Sharing Modeling, Simulation, and Analysis Tools	52
6 Summary of Crosscutting Findings and Recommendations	55
6.1 Findings.....	55
6.2 Recommendations.....	56
7 Glossary	59
7.1 Terms	59
7.2 Acronyms.....	61
8 References	65
9 Participants and Contributors	69
10 Acknowledgments	71

Cover Image: Front view of the semiconductor trackers for ATLAS, one of the four enormous detectors for the Large Hadron Collider at CERN, starting in 2007, built at NIKHEF, The Netherlands. (Credit Peter Ginter/NIKHEF)



1 Executive Summary

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. Participants in the review included facilities staff, program managers, and scientific experts from the offices of Basic Energy Sciences, Biological and Environmental Research, High Energy Physics, and Advanced Scientific Computing Research. Additional input came from previous workshop reports, as well as responses to a questionnaire provided by many facilities and science communities.

The review began with a series of talks from facilities operators and physical scientists who surveyed the enormous scientific advances that will be enabled by upcoming enhancements to experiments and detectors. As the presentations detailed, these advances will greatly benefit from solutions to a host of difficult data science challenges. Technical solutions to these challenges will foster new science understanding opportunities. Deeper understanding of Higgs Boson properties and its implications for the fundamental laws of nature will result from improved methods for managing and analyzing the enormous data sets generated by the Large Hadron Collider. Real-time analysis of very large sky survey data and retargeting of additional observational resources will enhance understanding of supernovas. Analysis of time-resolved experiments at DOE light sources, neutron facilities, genome sequencing facilities, and nanoscience centers will spawn new insights in biological and materials science. A deeper understanding of climate science will result from improvements in the management and interaction between large experimental and computational data sets.

Anticipated facility enhancements also will greatly increase data volume, velocity, and complexity. Although the resulting data will have greater scientific value, this information will challenge existing approaches to data collection, management, and analysis. One common concern was that current scientific data infrastructure will not manage this impending data growth.

As part of the meeting, review participants assembled into breakout sessions to discuss the key issues associated with three distinct aspects of the data challenge: 1) processing, 2) management, and 3) analysis. These discussions identified commonalities and differences among the needs of varied scientific communities. They also helped to articulate gaps between current approaches and future needs, as well as the research advances that will be required to close these gaps. Moreover, the review provided a rare opportunity for experts from across the Office of Science to learn about their collective expertise, challenges, and opportunities.

The review generated specific findings (further detailed in the body of this report).

1 FINDING 1: **The challenges associated with scientific data are diverse and often distinct from challenges in other data-intensive domains, such as web analytics and business intelligence.**

The volume and velocity of scientific data can be extremely high. Scientific data are precious and can be impossible or expensive to regenerate. Transparency and access to scientific data are important considerations. Tools and technologies developed for other applications will not be sufficient to address all of the data science needs encompassed within the Office of Science.

2 FINDING 2: **Research communities across the Office of Science have considerable expertise in the aspects of data science necessary for performing their science.**

However, the data science communities in different parts of the Office of Science are not fully aware of each other's capabilities and often do not coordinate their activities, which can lead to inefficiencies and missed scientific opportunities. Greater coordination and communication across the Office of Science—in headquarters and among researchers—would be beneficial.

3 FINDING 3: **Many Office of Science experimental facilities anticipate rapid growth in data volume, velocity, and complexity.**

These facilities require end-to-end systems that can automate, to a much greater extent than at present, the ingestion, analysis, and management of increasingly large and more complex data sets that can support much greater data rates. Many core needs are similar across different facilities, but detailed requirements can vary significantly. Rapid growth in data rates will necessitate new analysis techniques to enable real-time decision making, near-real-time data reduction, and offline analysis of large data sets. These needs will require advances in statistics, machine learning, visualization, and other related areas. Substantial progress will require mathematicians and computer scientists to work closely with domain scientists.

4 FINDING 4: **Currently, many scientific facilities expect users to manage their own data.**

This is particularly true of facilities that support a large number of diverse experiments, e.g., light sources, nanoscience facilities, and neutron sources. A greater degree of centralized support for data management, analysis, storage, and remote access would have numerous advantages, including helping to address the challenges of impending data growth, enhancing efficiency by reducing duplication of effort, and providing more consistent analysis and higher quality archival support—all of which would create new scientific opportunities and support open access.

5 FINDING 5: **There is an urgent need for standards and community application programming interfaces (APIs) for storing, annotating, and accessing scientific data.**

Development of standards and protocols for distributed data and service interoperability is essential. API standards will enable collaborations and facilitate extensibility, where similar customized services can be developed across science domains. Such standardization will facilitate data reuse and data integration from multiple experiments. It also will be needed to initiate activities toward providing facility-wide data services.

Based on these findings, this report offers several recommendations.

1 RECOMMENDATION 1: **The Office of Science should support multidisciplinary teams to conduct research and development needed to address DOE's unique data science challenges.**

Many of the data challenges confronting DOE will not be solved without Office of Science investment. These challenges can only be met by teams of computer scientists and mathematicians working closely with domain scientists and facilities personnel. The following areas are high priorities for investment:

- Flexible infrastructure for data management, curation, storage, and remote access that can be shared across communities
- Efficient methods for data reduction, storage, and access
- Scalable methods for data analysis, including statistics, machine learning, and visualization
- Techniques for combining data from multiple experiments
- Modeling capabilities to support the optimal design of data management systems
- Techniques for using simulations in support of experiments and employing experimental data to validate simulations
- Services that allow for low-cost, intuitive access to powerful data collecting, management, analysis, curation, and sharing capabilities.

2 RECOMMENDATION 2: **DOE science facilities should provide more centralized support for data management, storage, analysis, and access.**

DOE scientific facilities are used in diverse ways by a range of scientific communities. Many facilities require users to fully manage their own data. Commonly, all data are moved to a user's home institution for management and analysis. Already, this approach is inefficient and likely to be untenable in the future. Facility enhancements will dramatically increase data volume and complexity, resulting in data sets that are too big to move and too complex to analyze without assistance. Greater emphasis on scientific transparency will require this data to be more broadly accessible. And, new science undoubtedly will be discovered by making connections across and between experimental data sets from different users and facilities.

3 RECOMMENDATION 3: **The Office of Science should develop a cross-organizational strategic plan for data science.**

Because data science cuts across communities and is broader than any single component within the Office of Science, coordination is essential. Any plan should provide a framework for investment and prioritization with each office and identify dependencies between them. Topics that should be addressed include data-sharing policies, data curation standards, data science facilities and services, and sustainable software development and deployment. Such a plan would lead to improved efficiencies and scientific productivity.

4 RECOMMENDATION 4: **Mechanisms should be created to enhance communication among the scattered data science communities within the Office of Science.**

There will be significant benefits from exchanges of experience, best practices, perspectives, and current challenges.



2 Introduction


Today, a consensus view exists espousing that scientific exploration and discovery have four contributing aspects: 1) theory, 2) experiment, 3) simulation, and 4) data-intensive analysis. The first three are often referred to as the “three pillars of science,” while data-intensive analysis is referred to as a “fourth paradigm” [HTT09]. Now, it is commonplace for a single simulation or large-scale experiment/observation to generate terabyte data sets. Data rates and volumes are projected to accelerate even further, resulting in petabytes from a single simulation run or experiment/observation session. In addition, there can be a great deal of variety and complexity in scientific data sets, representing data in different structures and formats. It is becoming prohibitive for individual scientific user facilities to manage and analyze these data sets by themselves. Thus, it is time to invest in new approaches to large-scale data processing, management, and analysis that both address the key data challenges and apply across scientific domains. Substantial research and development work will be required to establish the concepts, methods, and tools that support these new approaches.

The purpose of this “crosscutting review” report is to identify areas of commonality in the data processing, management, and analysis needs among the user communities supported by U.S. Department of Energy (DOE) scientific user facilities. In the past, Advanced Scientific Computing Research (ASCR) program research predominantly focused on the management and analysis of simulation data¹. In this report, we focus on use cases of an experimental/observational nature. There are three reasons for this choice. First, because they are essential to validating simulation models, experimental/observational data are intertwined with simulation data. Moreover, simulations often

are used to guide experimental design and analysis. Second, management and analysis of experimental/observational data introduce scientific challenges that are distinct from those associated with simulation data. For example, verifying the correctness of data from experiments/observations requires techniques that identify outliers or noise. Third, the ability to compare large-scale simulation and experimental/observational data poses new challenges not apparent when data volumes were smaller.

This report stems from an April 2013 DOE requirements review that brought together domain scientists, mathematicians, and computer scientists from across the Office of Science and academia. They discussed relevant scientific drivers from three DOE offices, describing the operation and current data requirements, as well as projecting future data needs. The scientific drivers from the Office of High Energy Physics (HEP) include the Energy Frontier, Intensity Frontier, and Cosmic Frontier. Scientific drivers from the Office of Basic Energy Sciences (BES) are light sources, nanoscience centers, and neutron experimental facilities. The scientific drivers from the Office of Biological and Environmental Research (BER) include climate science, genome science, and environmental molecular science. While these science domains naturally vary in their data requirement details, the assembled group sought to find common crosscutting themes and challenges by employing three breakout sessions that each focused on identifying crosscutting computer science and mathematics requirements for a distinctive aspect of the scientific exploration process, including data processing, data management, and data analysis.

¹ http://science.energy.gov/~media/ascr/pdf/program-documents/docs/Crosscutting_grand_challenges.pdf



Herein, **data processing** refers to activities that must take place while data are collected from experiments/observations. These activities include data acquisition, data reduction, data transformation for subsequent analysis, data movement to remote sites (e.g., for data storage), workflows for multitask pipelines, and automatic collection of metadata and provenance regarding the data being collected. **Data management** refers to activities associated with storing, searching, and sharing data. These activities include: input/output (I/O) acceleration to storage systems, data retention techniques, tools for data sharing within and across communities, search tools for identifying subsets of interest, and tools that support data models for representing the domain view of the data. **Data analysis** refers to techniques and tools to extract knowledge from data, including methods and algorithms for enhancing data quality, various statistical and machine learning techniques, multi-resolution and multi-sensor analysis methods, and large-scale visualization techniques.

The remainder of this report is structured as follows: Section 3 presents a set of science use cases and includes recommendations of top-priority items. Section 4, which summarizes the findings from the three computer science and math breakout sessions, also includes high-priority recommendations in each area. The report concludes with a high-level summary of findings and recommendations distilled from the specific recommendations in each section.

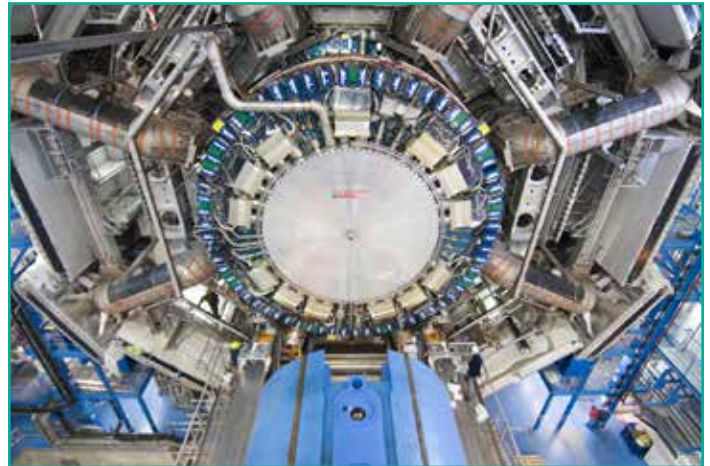
3 Scientific Drivers

3.1 High Energy Physics

3.1.1 The Energy Frontier

Located at the Large Hadron Collider (LHC)² in Geneva, Switzerland, two general-purpose detectors, designed to investigate a broad program of physics opportunities, are poised to expand knowledge at the Energy Frontier. The two detectors—Compact Muon Solenoid (CMS)³ and A Toroidal LHC Apparatus (ATLAS)⁴—are able to select and record high-energy proton-proton collisions and track the position, momentum, energy, and charge for each of the resulting particles from those collisions. Armed with that information, scientists can attempt to reconstruct what particles were generated by the collisions. By looking at ensembles of events, deeper insights into how the universe was created and the fundamental laws of nature, from the sub-atomic level to the more familiar macroscopic level, are expected. The analyses performed at the LHC typically encompass two varieties. One involves precision, where we intend to better understand the parameters of the Standard Model. The other involves searching for anomalies and possible clues that indicate the existence of new physical theories.

The recent LHC-generated observation involving what most people now believe to be the long sought-after Higgs Boson at 126 GeV has raised as many questions as it has answered, including: why is it so light? Does it imply a metastable universe, where the fundamental parameters are fine-tuned, or is there some undiscovered symmetry that explains it? If such symmetry exists, there are many particles yet to be directly detected at the LHC. One possible scenario of an underlying new symmetry, referred to as “supersymmetry,” includes a new particle that



View of ATLAS cavern side A while End Cap Calorimeter is being moved. (Courtesy CERN)

is stable, heavy, neutral, and not interacting electromagnetically. Such a particle, called the “neutralino,” would be a viable dark matter candidate, solving yet another mystery. Energy Frontier experiments are well positioned to pursue and offer a much deeper understanding of matter and the nature of space-time.

The small cross-sections of interesting physical interactions compared to backgrounds are at the heart of the computing challenges affecting LHC experiments. Discovery physics always is buried beneath a large, mundane, Standard Model background that is not easily modeled. The complexity of these collisions demands large, fine-grained detectors, leading to large recorded event sizes. Thus, full spectrum techniques are required to deal with the data, from specialized triggering hardware to massive amounts of commodity computing for storage and computation.

The LHC extends and will continue to push the boundaries for data-intensive science.

²<http://home.web.cern.ch/about/accelerators/large-hadron-collider>

³<http://cms.web.cern.ch/>

⁴<http://atlas.ch/>

Recording every collision would mean writing out almost a petabyte every second. While hardware triggers are employed to identify only the most energetic events, the LHC, nonetheless, already has acquired data sets encompassing many tens of petabytes of observational data and an equal amount of simulated data. By the mid-2020s, the LHC will be well on its way to dealing with exabytes of data.

Many aspects of future data requirements relate to the nature of the data. These issues cover topics such as data representations (e.g., event models) and data structures (e.g., data layout of events in storage, moving beyond file systems) designed to optimize data manipulation and analysis and data organization to enhance efficient selective data access (database design and indexing). Automatic collection of metadata from large-scale runs (e.g., workflow management) was another area of interest for all HEP research frontiers, while data archiving and curation (or knowledge preservation) also emerged as a significant crosscutting theme. As data volumes continue to grow, real-time monitoring of experimental and observational data was identified as another challenging requirement.

A central aspect of conducting data-intensive science is the manipulation, exploration, and analysis of data to extract scientific knowledge. This requires the development and usage of a host of tools and methods tailored to data-intensive applications: scalable and approximate algorithms, machine learning methods, experiment and simulation design, and advanced statistical techniques (regression; solution of inverse problems). Applications include anomaly detection, coverage of gaps in observed data sets, design of simulation campaigns, and experiment optimization. As data sets grow in size and complexity, uncertainty quantification and verification and

validation also become ever-increasing concerns to HEP scientists. Finally, the size and complexity of the data analysis chain have reached the point where workflow tools are essential for a large number of HEP science cases.

Recommendation: Develop tools that can support layout of data for efficient selective data access, automatic collection of metadata, and reliable data archiving.

Recommendation: Develop advanced scalable analysis methods, including advanced statistical methods and uncertainty quantification, powered by robust workflow technology.

3.1.2 The Intensity Frontier

The Intensity Frontier is characterized by experiments using intense particle beams and/or highly sensitive detectors to study rare processes more precisely and with more sensitivity than ever before. Neutrinos, though ubiquitous in the universe, are elusive and require intense beams and/or vast detectors to observe. Measurements of mass and other properties of neutrinos have profound consequences for understanding the evolution of the universe. Observations of rare processes that require exquisitely sensitive detectors and intense beams also explore high energies, providing an alternate, powerful window into the nature of fundamental interactions.

Current Intensity Frontier experiments collect 10,000s of files and hundreds of terabytes per year, requiring the same types of tools, software systems, and infrastructure as experiments that are generating petabytes per year, but are faced with the challenge of developing, supporting, and operating systems that are commensurately smaller relative to their data volumes.



The Far Hall (EH3) water pool with 3 Anti-neutrino detectors in place before the pool cover is installed.

The recent measurement of the neutrino mixing angle θ_{13} by the Daya Bay experiment⁵ was one of Science magazine's top 10 "Scientific Breakthroughs of 2012" and can be viewed as a typical current-generation Intensity Frontier experiment. Daya Bay detectors are located in China and generate ~400-500 gigabytes per day (24/7 year round), resulting in ~150 terabytes of raw data per year. The data are moved to compute facilities in Beijing and Berkeley (California) and stored, archived, and analyzed within minutes to hours of being collected. To analyze the entire data set, large-scale production analysis campaigns are conducted two to four times per year, while large-scale simulations using Geant4⁶ are conducted one to two times per year. The scale and complexity of the data movement and workflows require the same set of capabilities as the larger Energy Frontier experiments, although at a smaller scale.

Future Intensity Frontier experiments run the gamut from smaller experiments, similar in scale and complexity to the current generation, to larger experiments that will have requirements similar to those of current LHC experiments.

For example, the next-generation Belle II⁷ experiment, set to start operation in 2019, is expected to produce 30-50 petabytes/year of raw data. The raw data will be duplicated across two to three international sites, initially Japan's High Energy Accelerator Research Organization, known as KEK, and Pacific Northwest National Laboratory (PNNL). Reduced data will be further distributed across an international network of facilities for additional analysis [ADH+12].

For both present and future Intensity Frontier experiments, the range of I/O requirements relative to the central processing unit (CPU) requirements is particularly notable. Many experiments like Daya Bay have high I/O-to-CPU ratios, making them much more data intensive for a physical file system than analyses with heavier CPU requirements per byte of data.

Current Intensity Frontier experiments typically are significantly smaller than the large LHC experiments, both in data volume and rate. Nonetheless, these experiments also lack available sources to manage and analyze the resulting data and are critically deficient in manpower that can be applied to solve subsequent data-intensive science challenges. Consequently, the importance of broad efforts by HEP and ASCR to solve data-intensive science problems that can be easily adapted to and adopted by Intensity Frontier experiments cannot be overemphasized. Future Intensity Frontier experiments, such as Belle II, could reach levels of scale and complexity on par with leading Energy and Cosmic Frontier experiments and, therefore, share their more extended challenges and requirements.

⁵ <http://dayabay.ihep.ac.cn/twiki/bin/view/Public/> and <http://neutrino.physics.berkeley.edu/>

⁶ <http://geant4.cern.ch/>

⁷ <http://belle2.kek.jp/>

Recommendation: Develop tools for real-time analysis of entire data sets (either locally or remotely) and for collecting provenance information automatically from workflow processes.

3.1.3 The Cosmic Frontier

The Cosmic Frontier program focuses on the detection and mapping of galactic and extra-galactic sources of radiation and particles to reveal the fundamental nature of the universe. Current science thrusts within the Cosmic Frontier are investigations of dark matter and dark energy, high-energy cosmic and gamma rays, constraints on the neutrino sector, primordial fluctuations, and studies of the cosmic microwave background (CMB). The 2011 Nobel Prize in physics to Saul Perlmutter, Brian Schmidt, and Adam Riess for the “discovery of the accelerating expansion of the Universe through observations of distant supernovae,”⁸ marks the latest high point in Cosmic Frontier research.

From the perspective of data-intensive science, the Cosmic Frontier presents a number of challenges, ranging from real-time computing pipelines to large-scale analytics on massive data sets to the solution of computationally intensive inverse problems using the latest techniques from machine learning and applied statistics. The data sources include large-area cosmological sky surveys, particle detectors and CMB telescopes (in space and on the ground), and large-scale computations conducted on high-performance computing (HPC) platforms. Experiments, such as direct dark matter detectors, have data requirements that fall in roughly the same class as Intensity Frontier experiments and will not be discussed here in detail.



An artist's rendering of the proposed Large Synoptic Survey Telescope. The 8.4-meter LSST will use a special three-mirror design, creating an exceptionally wide field of view and will have the ability to survey the entire sky in only three nights. (Courtesy: LSST Corporation)

A motivating example for Cosmic Frontier requirements is the Large Synoptic Survey Telescope (LSST)⁹ project (construction start slated for 2014). The LSST will generate ~15 terabytes of raw data per night, amounting to about 100 million sources, of which about a million will be variable and announced (in near real time) as potential transients. The image data sets will be ~6 petabytes per year, representing a 20- to 40-fold increase in data volume and throughput over current data sets. To deal with this data flood, new scalable databases and analytics frameworks are required. As a data source for precision cosmological studies, LSST requires the development of a new inverse problem strategy that uses sophisticated end-to-end “forward model” simulations in conjunction with cutting-edge statistical approaches, such as likelihood-free inference.

⁸ http://www.nobelprize.org/nobel_prizes/physics/laureates/2011/press.html

⁹ <http://www.lsst.org/lsst/>

As with the Energy Frontier, detailed simulations of experimental observations are becoming a requirement for the Cosmic Frontier scientific program. This work includes large cosmological simulations and techniques to generate realistic synthetic sky catalogs so that a complete test of the observation-analysis-inference chain can be conducted before the experiment is undertaken. The high level of accuracy requirements (~1% or better) means that uncertainty quantification, verification, and validation are crucial issues that need to be addressed. The data sets from simulations already are at the ~5 petabyte scale or more. Therefore, exercising these requirements will provide an essential testing ground for the observational program.

Cosmic Frontier research areas overlap strongly with a number of ASCR interests in data processing (enabling, automating, and capturing the scientific process), management (optimization of the processes involved), and analytics (actual analysis of the science data stream). Other topics of broad common interest include data plans, future technologies tracking, software integration and sustainability issues, and developing a larger “data vision.” In particular, data processing covers topics such as data representations, provenance, schemas, workflows, and quality assessment. As exemplified by the LSST example, real-time data reduction issues also can play an important role. Data management topics relevant to the Cosmic Frontier include: indexing; organization of the data on storage; methods for selective access; and data sharing, both in bulk and in real time (e.g., LSST alerts). The analytics domain covers a broad scope, including sampling and experimental design, robustness to data

shortcomings (e.g., size, sampling, foregrounds, or noise), inverse problems for parameter estimation, approximate algorithms (e.g., balancing performance and error controls), end-to-end propagation of uncertainty, and collaborative visualization.

Recommendation: Develop data processes to ingest, classify, analyze, and manage data sets in real time.

Recommendation: Create scalable and flexible data management and tiered analysis systems that allow for rapid customization to meet the evolving needs of multiple Cosmic Frontier experiments.

3.2 Basic Energy Sciences

3.2.1 Data-intensive Computing for Light Sources

The six major BES light source facilities include: the National Synchrotron Light Source (NSLS)¹⁰ and National Synchrotron Light Source II (NSLS-II)¹¹ at Brookhaven National Laboratory (BNL), the Advanced Photon Source (APS)¹² at Argonne National Laboratory (ANL), the Advanced Light Source (ALS)¹³ at Lawrence Berkeley National Laboratory (LBNL), and the Linac Coherent Light Source (LCLS)¹⁴ and Stanford Synchrotron Radiation Lightsource (SSRL)¹⁵ at SLAC National Accelerator Laboratory (SLAC). Increased data volume and collection rates at these light sources stem from several factors, including a new generation of fast, two-dimensional (2-D) detectors; increased flux/brightness that allows shorter exposure times; robots and other automation that improve sample throughput; and greater interest in time-resolved *in situ* experiments. Two characteristic examples of

¹⁰ <http://www.bnl.gov/ps/nsls/About-NSLS.asp>

¹¹ <http://www.bnl.gov/ps/nsls2/about-NSLS-II.asp>

¹² <http://www.aps.anl.gov/>

¹³ <http://www-als.lbl.gov/>

¹⁴ https://portal.slac.stanford.edu/sites/lcls_public/Pages/Default.aspx

¹⁵ <http://www6.slac.stanford.edu/facilities/ssrl.aspx>



APS – Upgrade of superconducting iD

data rates and volumes are an APS tomography beamline [WDM+01] that can produce 150 terabytes of data in one day if its detectors run at maximum capacity (although this is much more than the current average daily volume) [DXF+12] and NSLS-II, which will produce an average of ~75 terabytes per day after only the first few years of operation (15 petabytes per year). The variety of data at light sources compounds the challenges posed by increasing data volumes and rates. APS has ~60 beamlines, and ALS has ~40. With 58 planned beamlines, NSLS-II currently is in an advanced construction phase that will be commissioned and operational by mid-2014. Light source experiments range from ptychography, microscopy, and tomography to photon correlation spectroscopy, as well as various forms of scattering and crystallography, among many others. Each beamline can accommodate a multitude of specific science applications, which multiplies the variety again. This diversity assures that, for many challenges, there is no “one-size-fits-all” solution. However, there are some common challenges in terms of data management, processing, and analysis. As such, common solutions developed in collaboration with ASCR could benefit all light sources.

Few light sources define facility-level standards for data management, sharing, or preservation, and few centralized facility-level resources are dedicated to these purposes (LCLS is the exception). Rather, data management, processing, and analysis are handled on a beamline-by-beamline basis, and the standards and available resources vary widely. As light sources consider changes to their standards, they must be responsive to user needs but also cognizant of the hierarchy of data standards, including those of funding agencies, facilities, journals, and user institutions.

A common “data management” method at light source beamlines is the “manually-copy-data-to-a-USB-hard-drive” method. At beamlines with high data rates, this approach is no longer feasible—too many hard drives would be required and data cannot be copied to them before a user's beam time ends. For these beamlines, there is no choice but to make substantial changes to the current data management approach. In addition, current methods and hardware for data processing and analysis are falling farther and farther behind data collection. Facilities acknowledge these challenges. However, they also have overarching concerns relating to the high cost of both compute hardware and staff that will be required to make changes and to the difficulty in training both staff and users in new systems and policies.

There is an increasing consensus that a beamline-by-beamline approach to responding to data challenges cannot support future data-sharing needs. LCLS and NSLS-II, the newest light sources, are the most oriented toward a centralized approach. An NSLS-II taskforce has been established to evaluate data requirements and provide recommendations for cost-effective data management approaches. Almost all facilities have pilot programs to provide tools and resources geared toward centralized data storage, processing, and analysis. In several cases, light sources are partnered with Office of

Science compute facilities, such as the National Energy Research Scientific Computing Center (NERSC). For example, a comment issued by researchers from one BES facility attending this workshop was that “the most appropriate role for [us] in sharing and preserving data is to partner with Office of Science Leadership Compute Facilities, such as NERSC.”

For some beamlines, because of budgetary constraints, there is no viable alternative to a centralized approach that relies on supercomputing facilities for handling current data volumes and rates, and many more beamlines will have this problem in coming years. Furthermore, providing a centralized platform for data storage and sharing has other advantages, such as facilitating collaborations between light source users. If methods for metadata storage can be developed, light source data are sufficiently rich that, in many cases, groups beyond the initial users might be interested in analyzing them. Data accessibility will lead to cross-checking of results and software, improved comparison between experiment and simulation, and afford more and better software development and access to developed software. Lack of access to both hardware and software for processing and analysis, not just data management and storage, is a key bottleneck for many users. Because much of the current software is available only at workstations or small clusters located at the beamlines, it is difficult for users to take advantage of these resources outside of their beam time. Providing user-friendly tools for data-intensive analysis on data sets residing at compute facilities, for example, through web interfaces, is a component of pilot projects at multiple light sources. Staffers from another facility in attendance summed up the

rationale for this approach: “Users should have easy access to their raw data, metadata, data reduction, and analysis code... Easily accessible data management and analysis capabilities represent the most effective approach to enhance user productivity.”

Recommendation: Develop scalable, robust, and reliable multi-facility capabilities for data management, analysis, archiving, and remote access that support the diversity of light source science experiments.

3.2.2 Emerging Data Challenges at the Nanoscience Centers

The DOE experimental nanoscience program encompasses the following BES user facilities:

- Five Nanoscale Science Research Centers (NSRCs):
 - The Molecular Foundry (TMF)¹⁶ at LBNL
 - The Center for Functional Nanomaterials (CFN)¹⁷ at BNL
 - The Center for Integrated Nanotechnologies (CINT)¹⁸ at Sandia National Laboratories (SNL) and Los Alamos National Laboratory (LANL)
 - The Center for Nanoscale Materials (CNM)¹⁹ at ANL
 - The Center for Nanophase Materials Sciences (CNMS)²⁰ at Oak Ridge National Laboratory (ORNL).



The Center for Functional Nanomaterials (CFN) at BNL.
(Courtesy: Brookhaven National Laboratory)

¹⁶ <http://foundry.lbl.gov/>

¹⁷ <http://www.bnl.gov/cfn/>

¹⁸ <http://cint.lanl.gov/>

¹⁹ <http://nano.anl.gov/>

²⁰ <http://www.cnms.ornl.gov/>

- Three Electron-Beam Microcharacterization Centers (EBMCs):
 - The National Center for Electron Microscopy (NCEM)²¹ at LBNL
 - The Electron Microscopy Center for Materials Research (EMC)²² at ANL
 - The Shared Research Equipment (ShaRE)²³ User Facility at ORNL.

This group of user facilities represents a wide spectrum of scientific disciplines. Given the range of imaging and spectroscopy techniques accessible to electron microscopes is common to multiple EBMCs, they are broadly similar. However each EBMC site/instrument has its own material/environment/resolution emphasis. While the NSRCs also have some electron microscope capabilities, they provide users with access to various synthesis laboratories (inorganic, organic, and biochemical/biological); top-down nanofabrication (e.g., lithography); imaging, spectroscopy, and manipulation; and theory and computation. Unique tools with growing data requirements include electron tomographic reconstruction, multimodal nanoscale *in situ* imaging, combinatorial synthesis facilitated by robotics, and computational surveys of materials properties and simulated characterization.

Currently, an array of data management policies exists across each of these facilities. Common to all is the sense that the ultimate responsibility for data resides with the users. Users typically arrive at a given facility with their own portable mass storage devices or with the expectation that they will transfer data to their home institutions before their projects conclude. Currently, data volumes are small (rarely terabytes). Most facilities rely on in-house file servers to store user data for the short term—perhaps the duration of a given user project, which may last about one year. In addition, access to data from outside of

the facility often is limited by the local policy (or culture) of data sharing and access. Moreover, useful analysis tools may only be available on site and require active facility support staff to enable remote analysis. Data sharing requires development. Currently, users take advantage of cloud-based solutions, such as Dropbox or Google Drive, but this type of solution will not scale.

Looking ahead, the near future (perhaps a five-year horizon) surely will be dominated by increased data demands from higher-resolution and higher-speed imaging; multimodal spectroscopy (increased dimensionality in measured data); combinatorial/robotic synthesis and characterization; and online analysis tools and computation, potentially coupled to local HPC resources. It seems apparent to almost all facilities that increases in file storage and data management infrastructure and support will require significant investments in equipment, human resources, and intellectual support for the various problems that arise concomitantly with larger data.

Efforts are underway to expand data sharing within collaborative teams. However, there are practical, cultural challenges in achieving data sharing among independent research groups with differing scientific goals. It may be that there is no close analog to other community-driven efforts to collect and share data, such as in astronomy, cosmology, or high-energy physics. However, facility staff at these experimental nanoscience institutions see commonality across many projects, recognize the need for improved data acquisition and analysis rates, and acknowledge the need to enable users to spend more time off site while remaining connected to their data and analyses through user-friendly interfaces. Ultimately, this would enable the facility to focus on serving more users more efficiently. Various efforts related to developing robust online (and/or remote) access and analyses have begun and could benefit from lessons learned

²¹ <http://ncem.lbl.gov/>

²² <http://www.msd.anl.gov/groups/emc/>

²³ <http://www.ornl.gov/sci/share/>

within other science communities. Furthermore, sharing data across user projects or between facilities through common data-conversion tools may enable unique analysis at one facility to be applied to a wider range of data common among more than one facility—especially in the facilities with electron microscopes. These large databases will provide a rich source of information for data-mining techniques whose aim is either to improve measurement accuracy or expedite the search for desirable materials, as exemplified by the Materials Genome Initiative.

An example of the type of development that would meet user needs, while simultaneously presenting problems of interest to ASCR, currently is underway at TMF (LBNL). WebXS is an online tool that simulates and interprets X-ray spectroscopy of three-dimensional (3-D) atomic models or trajectories, which initiates, analyzes, and visualizes first-principles simulations on HPC resources, such as NERSC. To interface an HTML5/JavaScript user interface with an HPC resource, WebXS exploits the NEWT application programming interface (API) and web service. Core-level absorption spectroscopy reveals element-specific information on local electronic structure (e.g., chemical bonding and coordination, oxidation state) and may be accessed using electron energy loss in transmission electron microscopes or, more commonly, using X-rays generated with table-top high-harmonic generation or at synchrotrons and (more recently) X-ray free-electron lasers (XFELs). When applied to real materials, the spectra contain a wealth of information, but their interpretation often is quite difficult without associated theoretical modeling or simulation. For a system of ~100 atoms, 10 gigabytes of data per atom could be generated, which can easily grow to a total size of ~100 terabytes when sampling molecular dynamics trajectories.



The SNAP Diffractometer at the Spallation Neutron Source
(Courtesy: Oak Ridge National Laboratory)

With Office of Science guidance, these DOE-BES user facilities can establish improved and more uniform policies with respect to data management, sharing, and preservation. Continued investment in various projects aimed at increasing data generation and analysis and improved user access will serve both to elevate current user community expectations and increase their data needs by making data collection and analyses easier. Ultimately, it seems highly likely that lessons being learned right now at other BES user facilities will be applicable to the NSRCs and EBMCs when they establish greater data needs.

Recommendation: Develop data conversion, data management, and remote data access tools across all experiments to enable data sharing between facilities.

Recommendation: Establish uniform policies for data management, sharing, and preservation.

3.2.3 Data-intensive Neutron Facilities

This section summarizes the needs and challenges for the three DOE-BES Neutron Scattering Facilities: the Spallation Neutron Source (SNS)²⁴ and High Flux Isotope Reactor (HFIR)²⁵ at ORNL and the Lujan Neutron Scattering Center²⁶ at LANL. Spallation neutron

²⁴ <http://neutrons.ornl.gov/facilities/SNS/>

²⁵ <http://neutrons.ornl.gov/facilities/HFIR/>

²⁶ <http://lansce.lanl.gov/lujan/>

sources (SNS and Lujan Center) generally produce larger data sets using time-of-flight of neutrons as an additional dimension compared to reactor-based sources (HFIR). In terms of data volume and velocity, the SNS will be capable of producing in excess of 150 terabytes of raw data each day of operation, assuming a full instrument and detector build of the facility. In addition to volume and velocity, the variety of instruments; detector types; and, most importantly, types of experiments across all neutron facilities are making a “one-size-fits-all” solution to data challenges impossible or at least impractical. Looking forward, a variety of experiments will extend across these institutions as researchers employ multiple DOE-BES user facilities to unlock scientific secrets.

Currently, experimental data files collected at the SNS are catalogued, stored, and replicated on the large-scale High Performance Storage System (HPSS) data system at the National Center for Computational Science (NCCS) at ORNL. On some beamlines, data are automatically reduced to, for example, a powder diffraction pattern. Data access is restricted to members of an individual team associated with the experiment. Users have access to the raw data, as well as reduced and processed data, through analysis workstations, a data portal, and within the data reduction framework Mantid²⁷. These resources are available to SNS users both on- and off-site. Data collected at HFIR and the Lujan Center currently are not centrally catalogued, although catalogues exist at individual beamlines. Offsite access to data from these facilities usually is done by contacting beamline staff directly. User agreements between the user’s home institutions and the user facility define ownership of the collected data.

One major challenge users encounter during their experiments on high-data-rate instruments at SNS is slow data reduction times, which

effectively disconnects the user from the experiment currently going on. The ADARA, or Accelerating Data Acquisition, Reduction and Analysis at the SNS, project implemented as a prototype on HYSPEC (hybrid spectrometer) is starting to address the issue and will provide live data viewing and access on all SNS beamlines over the next two years. This work will build the foundation for enabling automatic feedback between data analysis and instrument control.

A similar challenge is bringing advanced modeling and simulation techniques closer to the broad community by lowering the barriers to access HPC resources and providing hardened production-level analysis and modeling tools. Given the variety of experiments, these solutions will depend strongly on the field of science related to the experiment. In general, the science productivity of user facilities will be only as good as the weakest link in the data pipeline. Users often write their own code or use prototype community programs to analyze or model their neutron (and/or X-ray) data. A mechanism to move relevant scientific analysis and modeling software from a prototype developed as part of a specific scientific project to production level and continuous maintenance is urgently needed.

As complex science problems increasingly will require experimental data from multiple instruments and/or facilities, it will be essential to provide easy access and sharing of data and tools across DOE-BES facilities and beyond. This effort also will require suitable data policies across facilities to allow sharing and access by the original experimenter team, as well as the scientific community at large.

Recommendation: Develop tools for *in situ* data reduction at beamlines that can apply to multiple facilities.

Recommendation: Develop tools that aid in sharing and maintaining relevant scientific analysis and modeling software at production level.

²⁷ http://www.mantidproject.org/Main_Page

Recommendation: Develop common data access and analysis tools that support science across neutron and light sources, as well as joint data standards across facilities, to allow sharing by the scientific community.

3.3 Biological and Environmental Research

3.3.1 Climate Science Challenges and Motivation

Climate science is a prominent example of a discipline where scientific progress is critically dependent on the availability of a reliable infrastructure for managing, accessing, integrating, and comparing large quantities of heterogeneous data on a global scale. It involves an inherently collaborative and multidisciplinary effort that requires sophisticated observation and modeling of the physical processes and exchange mechanisms between multiple Earth realms (atmosphere, land, ocean, and sea ice). These models have been developed based on results from and are evaluated through comparison with observational measurement data from various sources—collected at different scales with different observational methods—possibly acquired over long periods of time.

The climate science community has a tradition of operating long-term observational measurement campaigns to track climatic conditions and build “climatologies” (multi-year records) that support in-depth studies of climatic process drivers. The results of these studies are vital to improve the understanding and representation of climatic processes in climate and Earth system models and resolve the uncertainties in such models toward the development of sustainable solutions for the nation's energy and environmental challenges. DOE-BER's Atmospheric Radiation



ARM Radar Van. (Courtesy: Pacific Northwest National Laboratory)

Measurement (ARM)²⁸ Climate Research Facility is focused on providing detailed and accurate descriptions of the Earth's atmosphere in diverse climate regimes via measurements from strategically located *in situ* and remote-sensing observatories. ARM collects several thousand data streams 24/7, which it composes into value-added products (VAPs) customized for the specific needs of the different atmospheric research communities it serves. Next to model development support, these data products also are used for model evaluation and validation. Increasingly, these VAPs are not built solely on ARM data but combined results from many similar measurement facilities. The latest strategic plans from BER²⁹ envision an increased need for integrative data products across scales and disciplines and an infrastructure that supports their generation and use in support of developing predictive modeling capabilities of complex, multiscale, coupled, and biologically based environmental systems behavior.

For the past several decades, the climate community has worked on concerted, worldwide modeling activities led by the Working Group on Coupled Modeling (WGCM), sponsored by the World Climate Research

²⁸ <http://www.arm.gov/>

²⁹ http://science.energy.gov/~media/ber/berac/pdf/20130221/BERACVirtualLaboratory_Feb-18-2013.pdf

Program (WCRP), leading to successive reports by the International Panel on Climate Change (IPCC). Currently, the fifth assessment (IPCC-AR5) is underway (due out in late 2013). These activities involve tens of modeling groups in as many countries, running the same prescribed set of climate change scenarios on the most advanced supercomputers and producing several petabytes of output, containing hundreds of physical variables that span tens and hundreds of years. These data sets generated by climate models are held at distributed locations around the globe, but they must be discovered, downloaded, and analyzed as if they are stored in a single archive with efficient and reliable access mechanisms that can span political and institutional boundaries. The same infrastructure also must allow scientists to access and compare observational data sets from multiple sources, including, for example, Earth Observing System (EOS) satellites and ARM sites. These observations, often collected and made available in real time or near real time, typically are stored in different formats and must be post-processed to be converted into a format that affords easy comparison with models. The need for providing data products, as well as VAPs, on demand adds another dimension to the necessary capabilities. Finally, science results must be applied at multiple scales (global, regional, and local) and made available to different communities (scientists, policy makers, instructors, farmers, industry, etc.). Because of its high visibility and direct impact on political decisions that govern human activities, the end-to-end scientific investigation must be completely transparent, collaborative, and reproducible. Scientists must be provided with the environment and tools for exchanging ideas and verifying results with colleagues in different time zones, investigating metadata, tracking provenance, annotating results, and collaborating in developing analysis applications and algorithms.

Computing services, supported by key climate modeling and data centers such as those hosted at DOE's Leadership Computing Facilities at ANL, ORNL, and NERSC, provide the climate community, in particular, with HPC, clusters, short- and long-term storage, networking, and coordinated software infrastructure resources that are spread throughout the climate federation. In contrast, observational climate facilities usually operate their own specific data and computing infrastructures. In addition to these capabilities, the community must rely on multiple levels of service to effectively produce, analyze, and manage distributed climate data from many sources:

- **Domain-specific Distributed Data Services:** Captures the set of unique requirements and needed services for each unique climate project.
- **Common Data Services:** Shared across all climate projects, such as movement, curation, discovery, exploration, analysis, etc.
- **Data Systems Software Layers:** Includes lower layers of software services, such as metadata, directory structures, provenance, and workflow.
- **Data System Hardware:** Includes HPCs, clusters, clouds, and large storage for modeling, *in situ* data analysis, and post-hoc large-scale data analysis of observational and computational results. This also includes in-transit processing to enable extreme-scale climate analysis.
- **Networks:** Binds the collection of disparate hardware, networks, and software resources for community use. Networks also are necessary to replicate and move large data holdings at storage facilities and federate connectivity. Energy Sciences Network's (ESnet) 100-gigabit network is of particular interest.

If DOE means to optimize its data investments, it must ensure that a common open architecture is in place and a significant fraction of that architecture is shared among the different climate activities rather than having specific domain architecture for each project.

Recommendation: Develop a reference model and supporting API standards to enable collaborations and facilitate extensibility, where similar, customized services can be developed across science domains.

Recommendation: Develop community-established standards and protocols for distributed data and service interoperability of independently developed data systems and services.

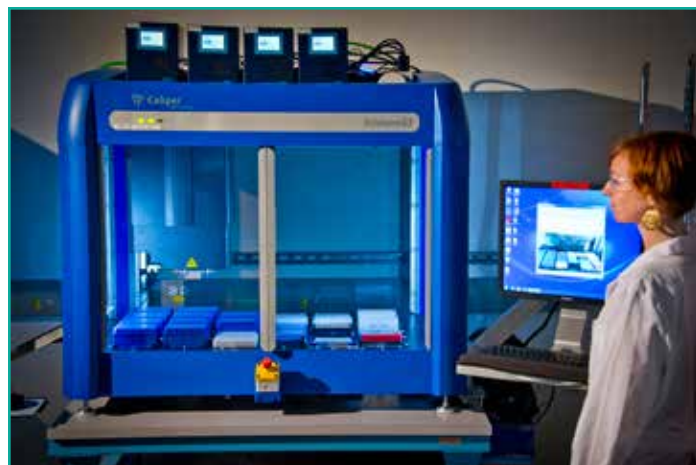
Recommendation: Develop effective parallel algorithms for analysis, integration, and comparison of heterogeneous observational and modeling results.

3.3.2 Computational and Data Challenges for High-throughput Genome Science

This section contains two parts, representing large-scale activities in Genome Science: the Joint Genome Institute (JGI)³⁰ and the Systems Biology Knowledgebase (KBase)³¹ project. JGI is a facility with distinct operational requirements, while KBase is a computational and data architecture framework for predictive biology with unique analytical requirements.

The Joint Genome Institute

JGI is DOE's sole production genome science facility. In 2012, JGI served nearly 1,000 users and sequenced 53 trillion nucleoside bases. In fiscal year (FY) 2013, it will sequence more than 72 trillion bases. JGI's stated objective "is



Caliper ScicloneG3 robot at The Joint Genome Institute
(Courtesy: JGI)

to couple the generation of sequence data with the development of new large-scale experimental and computational capabilities to functionally annotate DNA sequences, thereby narrowing the gap between the generation and interpretation of sequence data."³²

Since 2005, the genome science community, JGI included, has benefitted from exponential decreases in the cost per sequenced base as second-generation sequencing technologies (short-read, whole-genome shotgun sequencing methods) have matured. The greater than five times per year decreases in cost per base from 2005 to 2011 have paired with exponential increases in the sequence data collected. The expansion rate of sequence data has far outpaced Moore's law, leading to challenges working with all of these new data, particularly in reducing and analyzing the data. In 2010, to mitigate the issues of scale that these new data have generated, JGI partnered with NERSC to integrate HPC into their data analysis pipelines.

³⁰ <http://www.jgi.doe.gov/>

³¹ <http://kbase.science.energy.gov/>

³² <http://www.jgi.doe.gov/whoweare/10-Year-JGI-Strategic-Vision.pdf>

For the next one to three years, the expectation is that JGI's sequencing rate will remain approximately constant at ~70 TBase/year due to maturation of second-generation sequencing technologies; budget limitations; and some resource redirection to "third-generation" technologies, such as Pacific Biosciences' single molecule real time (SMRT) sequencing. Presently, these newer technologies produce lower total data volumes than the second generation, but they may experience higher data growth rates in the future (two to six years forward) as the technology matures. In the meantime, concurrent use of SMRT strongly complements JGI's short-read production sequencing capabilities because SMRT generates much longer reads that can serve as a scaffold to which the higher-quality second-generation short reads can be mapped. In addition, these newer technologies enable new and high-throughput experiments, such as concurrent sequencing and methylation-state analysis, which will greatly enhance biological insight but will concomitantly increase computational expense, as multiple data sets will be needed for simultaneous consideration.

The experimental data generated by JGI and transferred to NERSC (raw data, calculated sequence, quality scores) presently can be as high as 5 terabytes/day (68 megabytes/s). Data transfer or storage is not the major challenge in working with these data. Instead, it is reducing and analyzing the data. The two principal analyses where most computational effort is spent are assembly of the short reads and functional annotation of the finished sequence. Both of these analyses yield products of intense biological interest: an assembled genome and attribution of genes and other features of interest, respectively. With the integration of SMRT data, the assembly problem (~23% of JGI's computational effort) may be simplified in the future. The annotation problem, on the other hand, is far more challenging because each newly sequenced genome must be analyzed

in the context of many other genomes. Thus, the computational cost of analysis increases exponentially.

The critical algorithm used for comparative analysis of genetic sequences is the Smith-Waterman pairwise local alignment algorithm. This algorithm compares the sequences while applying a mathematical model of evolutionary constraints on permissible exchange rates of the bases to determine equivalent sequences (thus, direct text comparison is not feasible). A number of improvements have been devised, including the basic local alignment search tool, or BLAST, algorithm that implements approximate matches of a query sequence to a reference database of sequences, identifying likely targets for detailed alignment. In other alignment software (BLAT, USEARCH), additional heuristics have been implemented to gain further improvements. A major issue with pairwise sequence alignment is that the rate of I/O as the calculation is scaled up is insufficient to saturate the CPU. At JGI, one common analysis building on BLAST (or similar programs) is a gene-clustering analysis, where each identified gene is aligned to every other gene to obtain dissimilarity scores in a dense matrix to feed a hierarchical clustering algorithm. This method is accurate but expensive. There are new approaches (UCLUST), where a reference set of genes is taken to represent the clusters then target genes are iteratively clustered into these seeds. This approach is much faster, but it also has several limitations that may restrict its application (evaluation order dependencies, high-identity requirement within clusters).

The overall data size of genomics data *per experiment* is rather small compared to a high-energy physics experiment. However, these data are challenging to reduce and analyze because 1) a genome is inherently a string of characters that cannot be further reduced once a high-quality assembly has been constructed and 2) the annotation of a genome

is an iterative process, where each genome must be compared to many other genomes simultaneously. Improvements in file systems and I/O architectures will be immensely helpful in delivering these data fast enough to the CPU for analysis. Still, development of new algorithms to aid in comparative analysis will be critical.

Recommendation: Development of fast I/O techniques to leadership-class computers and new algorithms for dealing with exponential growth in the complexity of comparative analysis.

Computational and Data Challenges of DOE Systems Biology Knowledgebase

DOE's KBase is an ambitious effort to accelerate understanding of microbes, microbial communities, and plants and advance the foundational knowledge underlying biological approaches to producing biofuels, sequestering carbon, and cleaning up contaminated environments. KBase seeks to exploit the deep molecular knowledge of genetic functions and leverage the power of models to predict how diverse cells create organismal and ecological behavior driven by their genomic programs and determine which cellular interventions can produce a desired biological outcome. The overarching objective is an extensible model-based framework to support predictive biology, ranging from genomes to phenotypic characteristics and behaviors of microbes, plants, and communities.

To achieve these objectives, KBase must design and implement an advanced scalable computational infrastructure that provides an extensible tool set to facilitate predictive biology by allowing researchers to quickly and easily map genotypes to molecular, organismal, and ecological phenotypes. The key goals of the KBase infrastructure design are to:

1. Implement a system that provides users nimble access to the excellent, distributed HPC infrastructure of DOE laboratories.

2. Enable users to perform extremely sophisticated computation on large data sets efficiently in a unified framework where users can add data and algorithms to the system and service, which then would become available to other researchers quickly and in an integrated fashion.
3. Create metrics to compare the efficacy of state-of-the-art algorithms and assess the quality of user-uploaded data.
4. Develop a resource where users can formally share the process and thought by which they jointly arrive at their conclusions and discoveries.

The ultimate objective of KBase is to facilitate DOE biological research relating to bioenergy, carbon cycle, and the study of subsurface microbial communities.

For microbes, KBase will provide microbiologists with computational tools and infrastructure to map genotype to ecologically meaningful phenotype through experimentally supported models of cellular network function and enable the function to be harnessed for applications in energy production, remediation, and more. KBase will allow microbial researchers to infer a metabolic model fully reconciled with the data and visualized in concert with other available metabolic and RNA expression data. The framework will recommend candidate genes that fill critical holes in the metabolic model, thereby suggesting further experiments.

For plants, the science objective is to facilitate understanding of and modeling how genetic variations present in plant populations influence traits of interest. The notion that plants comprise multiple developmental, genetic, and functional regulatory systems underpins this modeling. Quantitative understanding of interactions within and between these systems would allow researchers to construct comprehensive dynamical models of plant cells; tissues; and, ultimately, entire organisms. The KBase plant

group's computational goal is to deploy a set of services that allows linking of gene targets from phenotype and genotype studies with co-expression, protein-protein interaction, and metabolic models and enables the inference of function of unknowns through statistical "guilt-by-association" methods.

For communities, KBase will support comparative analysis of metagenomes acquired over different spatial, temporal, or experimental scales. KBase also will help improve understanding of ecological traits and behavior, including how communities respond to their environment, evolve to deal with perturbations, and affect ecological and biogeochemical changes. KBase will enable researchers to use *in silico* experimentation to design the best sampling strategy. A new method for accurate and diagnostic assessment of the quality (or lack thereof) of metagenomic sequence read data will be used to evaluate experimental results. This framework will assess which changes to experimental protocols might generate higher-quality results. A second set of services will enable comparison and contrast of communities with similar alpha diversity or environments and will locate novel proteins for further study, as well as identify optimal candidates for screening.

KBase leverages JGI's existing high-performance sequencing, alignment, and annotation capabilities, as well as a broad spectrum of data types and sources across the microbial community and plant domains, and ties these data into a varied set of powerful computational tools that can analyze and simulate data to predict biological behavior, generate and test hypotheses, design new biological functions, and propose new experiments. The challenges span complex data analyses to understand individual and community behaviors; support workflow and provenance; and implement distributed data and computation infrastructure for searching, high-speed accessing, and processing.

Complex Data Analysis/Modeling Challenge:

KBase seeks to exploit the deep molecular knowledge of genetic function, discovered either in intact genomes or metagenomically mined from the Earth's genetic potential, to enable two key activities: 1) to predict how diverse cell types and their genetically idiosyncratic individuals, acting together, create organismal and ecological behavior driven by their genomic programs and 2) to predict which interventions, whether external or by modification of the cell's genome, can be made to achieve a desired biological outcome, such as increasing biofuel production or improving crop yield.

Sophisticated Workflow (Pipeline) Challenge:

Technology breakthroughs make it increasingly possible to measure an astounding array of features of biological systems and their constituent organisms, cells, and molecules. A single laboratory is capable of generating many terabytes of diverse data annually, ranging from DNA sequences to environmentally dependent phenotypic responses, molecular changes, population structural shifts, and environmental consequences. A single paper may report on terabytes of complex data analyzed using sophisticated algorithms with a number of free parameters to tune their performance. The results then are interpreted with the full complement of knowledge and insight from a deep and scholarly biological team. It is difficult to review the results of such papers thoroughly in a short amount of time, and it is nearly impossible to reproduce the results from the starting data. The KBase project aims to increase the transparency and reproducibility of such research by recording the process by which scientists use data and tools to draw conclusions.

Distribute Data and Computing Management, Access, and Integration Challenge:

The resources required to generate and analyze complex data (which often necessitates using high-performance compute clusters and large-

scale data storage) are beyond the means of most laboratories in the United States and worldwide. Many of the laboratories and user facilities that generate the largest data sets lack the resources to fully exploit, analyze, and integrate their own data sets. Thus, free and open access to data, the algorithms to analyze them, and the computational resources to make it feasible all are essential.

The KBase physical infrastructure is distributed across four laboratories, forming a suite of hardware resources that provides integrated and reliable capabilities. Each of the four KBase sites includes compute servers (1000 cores) and storage servers (about 0.5 petabytes). In addition, KBase operates the Magellan cloud with 7,500 cores for supporting production services and development of KBase capabilities. Rapid transfer of data between KBase sites is enabled by leveraging ESnet, which can transfer data at more than 90 gigabits per second. As more and more users choose KBase as their analysis and modeling platform, the KBase hardware architecture is designed to be scalable to support increasing data access throughput.

Recommendation: Develop search, discovery, and data accessing tools that facilitate information exploration and *in situ* analysis, comparison, and visualization.

Recommendation: Develop a pipeline and workflow system to support provenance and integrate distributed computing and storage resources.

Recommendation: Develop statistical methods to allow evaluating, comparing, and aggregating of multiple models and methods over data acquired with different spatial, temporal, or experimental scales and create hybrid ensembles for improved prediction and knowledge beyond a single pipeline with individual tools and data sets.

3.3.3 The Environmental Molecular Sciences Laboratory

The Environmental Molecular Sciences Laboratory (EMSL)³³, a national scientific user facility sponsored by DOE-BER and located at PNNL, provides world-class fundamental research capabilities for scientific discovery and the development of innovative solutions to the nation's environmental and energy production challenges. EMSL's distinctive focus on integrating computational and experimental capabilities, as well as collaborating among disciplines, yields a strong, synergistic scientific environment. Bringing together experts and state-of-the-art instruments critical to their research under one roof, EMSL has helped thousands of researchers use a multidisciplinary, collaborative approach to solve some of the most important national challenges in energy, environmental sciences, and human health. These challenges cover a wide range of research, including synthesis, characterization, theory and modeling, dynamical properties, and environmental testing. EMSL houses a collection of more than 100 state-of-the-art capabilities.



The Aberration-Corrected and Monochromated Scanning/Transmission Electron Microscope in EMSL's Quiet Wing. (Courtesy: PNNL)

³³ <http://www.emsl.pnl.gov/emslweb/>

Time	Instrument	Data Rates	Burst	Volume
Today	Conventional	100-1000 Images/day	single	
1-2 years	Environmental	1000 Images/sec	10 min	11-13 TB/day
2-3 years	Dynamic	1,000,000 Images/sec	10 sec	20 TB/burst

Table 1: Current and Projected Data Rates for Transmission Electron Microscopes

Time	Daily Data Size	LAN Transfer	WAN Transfer
Today	6.5 TB/day	5 TB/day	200 GB/month
2-5 years	20-40 TB/day	40 TB/day	600 TB/month
5+ years	100 TB/day	200 TB/day	3 PB/month

Table 2: EMSL's Data and Storage Needs

Specific Data Challenges

Data Volume and Rate Challenge: Among other experimental techniques, EMSL houses a range of state-of-the-art transmission electron microscopy (TEM) instruments. TEM is a fast-growing imaging method with ~600 instruments currently worldwide, increasing by ~50/year. TEM instruments usually produce a series of 2000 × 2000 pixel or 4000 × 4000 pixel images at increasingly higher rates due to new detector technologies (shown in Table 1).

Presently, the complete analysis and interpretation of one image can take up to six months. However, the goal is to gain a deeper understanding of physical, chemical, and biological processes by analyzing all of the generated images. Furthermore, science would greatly benefit from real-time data processing capabilities to enhance the quality of data taking and enable interactions with the sample based on real-time results.

Multi-modal Challenge: EMSL experiments can involve multiple instruments (multi-modal) that may consist of a number of experimental techniques and computational simulations. Depending on the science being studied, this will require new techniques and tools for data assimilation and integration to explore the joint results. Data assimilation combines a number of data sources for comparison, including numerical simulations and observational data, using statistical methods and applied mathematics techniques. Data integration collects disparate data sets for meta-analysis (methods for contrasting and combining results from different studies, etc.). This type of data integration is especially challenging for multiple scientific disciplines, where there are many different data types produced in these fields. Table 2 shows how EMSL's data and storage needs continue to expand.

Cost Challenge: EMSL currently hosts more than 100 diverse instruments—each with its own specific analysis requirements. Available resources do not permit development of independent analysis solutions for each instrument. Another important concern for user facilities is the cost associated with maintenance and support of software stacks. Potential solutions might include:

- Coordinating with the user facilities to generate open-source efforts with the larger community (semantic physical sciences, NWChem, etc.)
- Identifying reusable and generally applicable software components (analysis tools, etc.).

Cross-facilities Collaboration: In the near future, new data plans, sharing, and data policies will require facilities to coordinate as much as possible to have more consistent capabilities that enable data sharing. It is conceivable that an experiment could be done at a beamline (e.g., at APS) and the same sample also could be taken to a user facility at a different laboratory (e.g., EMSL). If their respective data plans are not congruent, this could cause issues for users to obtain all of their data.

Recommendation: Develop tools that facilitate real-time processing and analysis of the large volumes of data collected by instruments, such as transmission electron microscopes.

Recommendation: Develop tools for flexible assimilation and integration of data generated by multiple sources of experimental, simulation, and observational data information using statistical methods and applied mathematics techniques.

Recommendation: Develop reusable analysis solutions that can be easily customized, extended, supported, and maintained.



4 Crosscutting Computer Science and Mathematics Challenges

Three broad categories of data challenges were identified as part of this workshop: data processing, data management, and data analysis. Again, data processing refers to activities that must take place while data are collected from experiments/observations, while data management involves activities associated with storing, searching, and sharing data. Data analysis references the techniques and tools needed to extract knowledge from data.

4.1 Data Processing

This section reviews the data processing requirements for experimental and observational DOE HEP, BES, and BER scientific user facilities, including:

- Identifying crosscutting themes in data processing requirements
- Highlighting short-term opportunities
- Providing recommendations for future research.

For the purpose of this report, data processing encompasses data acquisition, data reduction, data transformation, data movement, workflows, and metadata/provenance as they pertain to data processing.

Science Drivers

The scientific user facilities across the DOE offices generally have one of three operating models:

- 1) single-science focus/few experiments/many users (e.g., cosmology, particle physics, ARM),
- 2) multi-science focus/many experiments/many users (e.g., synchrotron facilities), and
- 3) many sites/many users (e.g., Earth Systems Grid Federation). Facilities of the single-science focus, with a limited number of experiments and many users, tend to have high data rates and highly automated and optimized data



This data-acquisition system is mounted on the electronics deck of MISSE 5. The microprocessor board (0) controls nine "daughter" boards (1 to 9) that record data for the experiment.
(Courtesy: NASA)

processing due to their concentrated centralized support. Multi-science (many experiments, many users) support facilities tend to have lower data rates per instrument but a significant number of instruments (tens), all different and each supporting a diverse range of sciences. Data processing needs vary greatly from instrument to instrument. Thus, at present, customized and ad hoc solutions dominate. In the third operating model, federated data access across facilities has been established, and they are perceived as large data clearing houses of analyzed results.

In each case, data processing is part of the core operational business of the user facility and must maintain a level of quality that enables fault-free, continuous operation 24/7, potentially for many months without the opportunity for maintenance or upgrades. In particular, the fault-free, reliable operation of data processing is absolutely critical as the majority of experiments are "one-shot" opportunities to capture the required measurement. Many experiments can never be repeated due to the singularity of the event (e.g., climate measurement) or the difficulty and costs of recreating the sample (many investigative methods are destructive) and the experiment itself.

Against this background of often highly diverse data processing requirements and the need for production-level software and hardware

solutions, there exists the challenge of increasing data volumes, which stresses existing data processing systems close to the breaking point. Over the past few years, developments in detector technologies for experimental sciences have resulted in a dramatic increase in data rates. Developmental detectors now can produce up to approximately terabytes/s and are expected to rise to ~petabytes/s by 2020. While these detectors are not yet in production at user facilities, they are slated to appear within two to three years. In the meantime, data rates already are increasing in the facilities—at times exponentially. Table 3 offers an overview of exemplary data rate projections from different scientific user facilities and instruments.

At times, the data rates in Table 3 already may overwhelm the acquisition system's ability to write collected data to disk for further processing, resulting in the need to reduce data on the fly. The LHC experiments require a data reduction exceeding 90% before data are written to disk. However, the ability to capture all of the generated data generally is limited not only by physical limits but also available resources. Thus, while it might be possible to engineer a solution that would capture and process all data, the costs would be prohibitive for the user facility, particularly if it needs to develop solutions for many different instruments. As such, it is important not only to focus on optimizing the technical solutions when

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

*Data rate after 99% reduction in hardware data acquisition system.

Table 3: Exemplary Data Rates for Different DOE Scientific User Facilities

designing the data processing infrastructure, but also on managing development and operating costs. The following sections discuss specific crosscutting requirements, opportunities, and recommendations.

4.1.1 Data Acquisition

HEP, BES, and BER are experiencing data rate (velocity and volume) increases in both experimental and computational systems that exceed the systems' physical capacity to write the streaming data out to disk for later processing. Current solutions rely on aggressive data reduction (e.g., HEP LHC 99% reduction) in the data acquisition infrastructure (hardware and software) to meet the challenge [Lipeles12, Youngman12].

Recommendation: Investigate the applicability of exascale computing research in memory access, I/O, and file systems to optimize data acquisition platforms.

Recommendation: Pursue the design of generalized data collection systems from an end-to-end perspective with ASCR researchers (computer science and mathematics) involved from the earliest stages to produce designs that maximize openness and flexibility and limit costs.

4.1.2 Data Reduction

Data reduction usually is split into two distinctive steps: 1) true lossy compression (e.g., triggering) to determine what data to keep and 2) identification of artifacts designed to eliminate noise and protect critical features. At present, the community uses fixed data reduction processes, such as triggers [Lipeles12] or creation of monthly means, to reduce data volumes. However, these approaches lack the flexibility to adapt their reduction to the unfolding results, for example, by collecting more data for rare events and less where nothing special is happening. The community wants a new

paradigm for data reduction that is adaptable and intelligent. With the advent of such adaptive methods, it is crucial that provenance information be captured throughout the process to enable verification of results and reproducibility.

Recommendation: Develop mathematical, adaptive workflow, and artificial-intelligence-based approaches for adaptively reducing data. Solutions must be driven by results and perform under given time and resource constraints.

Recommendation: Develop accompanying metadata and provenance capture, which is essential for adaptive reduction.

4.1.3 Data Transformation

The term "data transformation" can indicate either *data reorganization* for the purpose of enabling more efficient access, integration, and analysis (layout, indexing, clustering, etc.) or *data conversion* from one representation to another (e.g., different coordinate system, format, or vocabulary) for data integration. In data processing, both forms must be addressed. The optimized design of systems and algorithms for data transformation against the backdrop of large volumes of streaming data is a particular concern for most scientific user facilities.

Recommendation: Leverage ASCR research expertise in optimized exascale data layout and access for fast processing to design effective data transformation systems.

Recommendation: Leverage ASCR performance modeling expertise to assess the utility of specific solutions.

Recommendation: Demonstrate a new type of capability around data, focused on defining components (and associated APIs), that can be used and combined in flexible ways, as well as optimized automatically and dynamically for expected, observed, and unanticipated access patterns.

4.1.4 Data Movement

In the data movement arena, the challenges vary significantly between the different user facility groups, driven by their different operating models. As noted in ESnet requirements workshop reports,³⁴ regular, large-scale data transfer between known sites works extremely well on the ESnet network infrastructure, as transfer pathways are well maintained and optimized. User facilities that already use ESnet to its full extent (e.g., HEP Energy and Intensity Frontier facilities) would like help in optimizing their network usage (as part of their data processing environment) with features such as: finding available network paths; provisioning; interacting with active networks; selecting best paths; co-scheduling of compute, networks, and storage; and end-to-end monitoring, troubleshooting, and optimization.

The ESnet reports also highlight that ad hoc data transfers of even moderately sized data sets are a significant challenge. Many facilities with a widely distributed user base continue to rely on hard drives to transfer user data to users' home institutions. These user facilities need the means to offer high-speed data transfer to their users, either to their home organization or an attached data processing facility. Thus, they require data transfer tools that are easy to use and set up, hide idiosyncratic environments (e.g., firewalls, slow networks), and can automate problem diagnosis and correction. Once those capabilities are in place, they also would benefit from introduction of the optimization methods already described.

Often, the data movement process encompasses other activities, such as data synchronization, data sharing, metadata extraction, metadata publication to registries and catalogs, and provenance recording. Users will greatly benefit from tools that automate

these activities, whether as part of the data movement process or more sophisticated workflows.

With the advent of more distributed, collaborative research infrastructures, it also will become important to be able to optimize data movement across the different facilities and among users.

Recommendation: Use ASCR capabilities (ESnet and research) to create an easy-to-use tool for ad hoc data transfer from user facilities to users' target environments.

Recommendation: Research methods for automated and convenient optimization of end-to-end, multi-step scientific processes that involve multiple resources, users and applications, time periods, and activities, such as movement, synchronization, sharing, and publication.

Recommendation: Promote the development and use of environments at DOE facilities and laboratories that are tailored to the needs of high-performance science applications, such as the Science DMZ concepts.³⁵

4.1.5 Workflows

Workflows are a critical component of the data processing environment, enabling reliable automation and speed, and commonly are the "glue" for integrating many tools and technologies into an automated system. Like any other languages, there are a range of choices for the users, which often are incompatible with one another. One major problem is that each workflow language typically is tightly integrated with a specific workflow engine. This integration makes it impossible for users to use one common language and mix different engines. For example, there are workflow languages that work well on a single computer system, but they are not well suited for highly distributed execution. Workflow languages often are created with certain optimizations in

³⁴ <https://www.es.net/about/science-requirements/reports>

³⁵ <http://fasterdata.es.net/science-dmz/>

mind and are difficult to reuse when different optimizations become necessary. The new exascale paradigm of “power-awareness” is a prime example. Typically, workflow languages are created with the thought that workflows are “static” when no user intervention is required. In high-rate data processing environments, they need to create much more adaptive analytical capabilities, which must be supported by equally flexible workflow systems that allow user intervention, adaptive changes to the analysis, and re-optimization during these dynamic situations. Workflows also need to store their provenance information for “reproducibility.” Finally, a new generation of workflow engines needs to be researched that allows for billions of processes to work in a dynamic situation where there is no central workflow coordinator. Initial attempts to solve these problems exist in the Swift parallel scripting engine [WFI+09], the Middleware for Data Intensive Computing (MeDICI) workflow framework [GCW+09], and the workflow engine inside of the Adaptable IO System (ADIOS) framework [LZL+10].

Recommendation: Develop a knowledge base of the key characteristics of existing workflow systems and map those characteristics to known requirements within the scientific user facility complex. Study existing workflow systems to identify lessons learned and relevance to data-aware workflows. Investigate techniques that account for human computer interface (HCI) issues to allow easy construction of dynamic workflows.

Recommendation: Research new workflow engines and languages that are semantically rich and allow interoperability. Develop a new set of workflow engines that can be used interchangeably in many environments, from use over the wide area network (WAN) to exascale machines.

Recommendation: Research and develop a new paradigm for composing and executing dynamic workflows.

4.1.6 Metadata and Provenance

As processes become more complex, it becomes crucial to capture provenance and metadata throughout all processes and across systems to ensure reproducibility and enable verification of results. Therefore, provenance needs to be captured from automated processes (workflows), manual processes, and external sources (documents, software repositories) to afford completeness and “establish an uninterrupted chain of custody” [SPK13]. Furthermore, it must be possible to continue provenance capture across different processing steps (e.g., from initial data analysis into further use of the data to produce derived products or as input for validation tasks). Provenance capture alone, however, is not enough. Given the amount and complex nature of information that will be captured, it is equally important to investigate solutions that allow both systems and people to explore and use the collected information in the course of their science.

Recommendation: Develop provenance at scale to capture and exploit this information in future investigations.

4.2 Data Management

The data management breakout session had strong representation from BES, BER, HEP, and ASCR. The group discussed six topics and identified two or three recommendations for each. The topics included: data movement (I/O), data sharing, data retention and curation, search and discovery, data storage, and data models and schema.



4.2.1 Data Movement (I/O)

Managing data movement (i.e., data I/O) is critical in many contexts relevant to data-intensive science. For example, wide-area data movement is an important enabler for many science collaborations within DOE. The session attendees expressed concern that while science teams are making effective use of WANs to accomplish science goals today, there is an expected increase in the number of teams using these resources, which could lead to contention for these resources and degraded capabilities. The international collaborations in many science experiments make distributed data placement critical to effective data access. It also was recognized that a range of hardware architectures are deployed to interface complex detectors to data movement, storage, and analysis resources.

It was further noted that not all tools and applications being used in DOE experimental data analysis have adopted I/O best practices. Obtaining the highest performance into and out of complex I/O systems can be a complicated process, involving significant software development and tuning.

The attendees also recognized a need for more general abilities to tap into data streams for the purpose of various data analyses (e.g., anomaly detection or summary views of experimental data while experiments are in progress).

Recommendation: Research should be undertaken to develop new methods for scheduling data movement over wide-area links, including capabilities for providing quality of service and cost estimates. Research is needed to understand the impact of different replication policies, data architectures, and subset access mechanisms. Research into simulation and modeling of these systems also is needed, as well as availability of relevant test beds to experiment with alternatives.

Recommendation: Techniques should be developed to allow analysis tools and workflows to tap into data streams from instruments and simulations. These techniques must be incorporated into tools used in DOE science activities.

Recommendation: Certain science teams have a great deal of expertise using these systems, and it could be advantageous to the Office of Science for this experience to be more widely shared. Greater interaction between HPC I/O experts and science teams leveraging HPC resources and storage systems for large-scale data analysis is encouraged with the goal of passing on best-practice techniques to science teams.

4.2.2 Data Sharing

There are enormous benefits to making data accessible and sharable. Indeed, there is a presidential mandate for data sharing [HR10]. However, there also are important technical and non-technical barriers to sharing. In terms of the non-technical challenges, one is that many domain scientists have a “small science” outlook: they want complete control of their data and are not interested in sharing. There also is the issue of data policies of the respective funding agencies, facilities, and journals. Data sharing also raises numerous technical challenges, including how to integrate data from disparate sources. A significant amount of metadata and provenance must be collected

for data to be useful beyond the original domain scientists, and these metadata, along with data from multiple instruments, all must be integrated. Such metadata collection and integration likely will require development of shared data models and ontologies or schemas [MSF+10], but it also will entail deployment of tools and a cultural change among scientists to create a digital record of all relevant information about a given experiment that can be linked with the appropriate raw data and instrument metadata. Incentives for sharing should be investigated, for example, automated linking of shared data with other data sources. Providing tools that are easy to use and sufficiently functional so they catch on will be a challenge. Managing data security (single sign-on, access control lists (ACLs), virtual organizations (VOs), etc.) is an additional technical challenge associated with sharing, which is addressed, for example, by Globus Online [ABC+12].

A number of BES facilities commented that for the data their users collect, they do not have the expertise to provide this kind of data sharing capability to the scientific community and the public. Still, they would like to partner with Office of Science computer scientists to make this possible.

Recommendation: To achieve data sharing, ASCR should lead development of standards, tools, and services for collecting, annotating, preserving, and sharing data.

Recommendation: Engage in international efforts for data policies and sharing, such as the Research Data Alliance.³⁶

4.2.3 Data Retention and Curation

The retention and curation of recorded, retrievable research data are critically important in the pursuit of scientific integrity. “Data retention” is defined herein as the data

management actions that must be followed to maintain persistent records for long-term reuse. To this end, the scientific community and facilities responsible for generating or collecting scientific simulation and/or observational data have a responsibility to store, annotate, record, and retain research data for an agreed-upon period of time. This time period may range from mere days to an indefinite period. Motivations for data retention and curation can include:

- Use and reuse of key scientific results by the wider community to aid new discoveries
- Inform planning and design of future experiments
- Replicating experiments to reproduce scientific results for verification of conclusions
- Responding to scientific questioning and challenges
- Establishing owner identification of records
- Model calibration and feedback.

Additional data retention concerns involve data formats, provenance, archival rules, permissible storage, access, version control, and security—all within the data life cycle domain. The international community has researched this topic extensively (e.g., the Digital Curation Center³⁷, *International Journal of Data Curation*).

In some cases, data retention is linked to applications and workflows used to generate derived variables. Data sets may need to be retained only as long as they remain scientifically viable, and the cost of regenerating them outweighs the cost of their retention and hardware/software storage maintenance. As an additional cost-savings measure, some science domains have implemented standard data formats, conventions, metadata, software, workflows, etc., in support of deriving variables and reducing the need for data retention.

³⁶ <http://rd-alliance.org/>

³⁷ <http://www.dcc.ac.uk/>

For example, due to limited storage space or the cost of computing, some experiments are carefully documented—including information on the raw data, post-processing algorithms, experimental methodology, statistical treatments, hardware, operating system and compiler software, results and conclusions, and the time and conditions of the run process—in place of retaining data.

Recommendation: Review existing standards and policies and adjust for DOE's present and future requirements.

Recommendation: Develop best-practice data management and curation policies and guidelines for DOE science projects informed by existing and new international standards.

Recommendation: Develop tools and services that assist facilities and user communities in low-cost, long-term, high-reliability, and sustainable data retention/curation efforts.

4.2.4 Search and Discovery

Once data sets have been annotated and saved to persistent storage, the analysis and discovery work begins. Today's typical post-processing workflows involve investigating data with scientific data analysis tools, either custom or general purpose. Two general activities on the data are identified herein: 1) *search*, where a specific, defined query is made across data, and 2) *discovery*, where a more free-form exploration of the data is undertaken without a specific goal in mind. In the future, these two forms of data exploration should be more closely integrated, so scientists are free to ask whatever questions they can think of and are not fighting severe constraints within the tools. The more time domain scientists spend thinking abstractly about their science, instead of wrestling with tools, the better the science that can be done.

There are various levels of support for searching across the sciences. Usually, these are divided into communities that use databases and those that use file-based permanent storage. Because databases trivially support queries, some communities routinely use database queries for search. This typically takes the form of searching across the metadata that summarize a data set rather than searching across the actual data. Communities that rely on file-based storage often do not use search tools, relying instead on semantics built into file paths for simple searches (by user, date, etc.). There are examples of searching across file-based results. However, this method quickly runs into scaling problems if naïve approaches are used.

If search can be more fully supported for large-scale experimental and simulation data, there is prime opportunity for impact across domains. Notably, domain experts were not familiar with this way of thinking, so it is possible that supporting more powerful and abstract searches across large scientific data could significantly impact science results.

Metadata searches barely scratch the surface of what is possible with query-driven analysis. More interesting, and perhaps useful, queries can be performed by searching for features across many sets of results. This will require research in several areas, including:

- Scalable indexing of metadata and data
- Indexing in context of federated data
- Technologies for complex, richer queries
 - Semantic searching (either graph-based or complex queries)
 - Feature-based searches (including research into feature extraction algorithms)
 - Example-based searches
- Compression techniques and compact data representations that retain qualities of the source data sets

- System architectures for request-driven queries so that requested features might only be computed or extracted as needed.

Recommendation: Support research into searches across federated databases, including scalable indexing, complex queries, and request-driven queries.

Recommendation: Quantify the science impacts that more useable and powerful data searches of large experimental results can provide.

4.2.5 Storage

Data storage is a critical component of any data-intensive science initiative. For many DOE facilities and research programs, a scalable data storage infrastructure that supports a variety of data representations is the first need.

In the breakout sessions, domain scientists indicated they are facing a proliferation of heterogeneous storage technologies, formats, and data models. The storage technologies ranged from archival storage (tape), traditional network attached file systems, parallel file systems, and Structured Query Language (SQL) databases to NoSQL storage technologies. This heterogeneity in storage will remain and become even more prevalent in the future.

Storage formats ranged from simple files—standardized scientific data formats, such as Network Common Data Form (NetCDF) and Hierarchical Data Format 5 (HDF5), and serialized data objects, such as ROOT³⁸—to more rigorous data models with standardized metadata representation, such as NetCDF-CF³⁹ and NeXus⁴⁰. As described in Section 4.2.6, scientists require improved data models that support a broader range of scientific and engineering data sets.

In addition to heterogeneity, data-intensive science programs are increasingly reliant on federated data storage that may span multiple geographies for improved data availability, the need to collocate storage with local resources, and to handle data capacities that no single site can support. Scientists using these federated data storage systems require a global view of their data, the ability to access and manipulate their data as if it were local.

Recommendation: Develop a distributed data storage infrastructure that is accessible across all Office of Science facilities to support science teams that are increasingly reliant on the ability to access and process data across multiple facilities.

Recommendation: Provide science teams with well-supported tools for data movement/staging between experiments and centralized storage and assistance in the adoption and integration of existing storage tools and technologies, such as databases and scalable file systems.

Recommendation: Undertake research to develop new methods to support a global view of federated data storage and understand the tradeoffs between shifting data between data resources versus accessing remote data directly.

Recommendation: Conduct research to explore intelligent automation of caching data in heterogeneous, distributed data hierarchies.

4.2.6 Data Models and Schema

To handle, manipulate, query, and mine complex data, advanced data models and schemas must be created to enable the development of scalable and efficient functions to represent, manage, and analyze data. As the data sizes scale, so does the complexity of data. During the breakout sessions, the following increases in data analysis complexity were discussed:

³⁸ <http://root.cern.ch/drupal/content/root-files-1>

³⁹ <http://cf-pcmdi.llnl.gov/>

⁴⁰ http://wiki.nexusformat.org/Main_Page

- Data will have higher numbers of dimensions as scalability in compute performance affords more sophisticated models and ensembles.
- Data points potentially will be a combination of structured and unstructured data.
- Data often will exist at multiple scales and multiple resolutions.
- Observational data will need to be combined with simulation data for real-time feedback and post-processing.

For such complex data, traditional methods for developing data models are unlikely to work or scale. Furthermore, simple, flexible, and understandable schemas are important for end users. Techniques that do not easily incorporate domain knowledge or depend on a particular partitioning of data are unlikely to capture sufficient knowledge or be scalable, particularly for data analysis, data fusion, and integration.

Recommendation: Research and development of scalable and flexible data models for Office of Science problems are needed. These models would provide the foundation for implementation of storage, data analysis, and data management.

Recommendation: Research and development of representation and schemas that enable the specification of advanced data types, relationships, and efficient storage, as well as implementations via a variety of infrastructure software.

Recommendation: Development of methods that may represent different motifs, scenarios, flows, and types of analysis on top of simple and highly scalable infrastructure may be attractive.

4.3 Data Analysis

This section describes the data analysis needs of scientists using BER, BES, and HEP experimental and observational facilities. Herein, the term



"data analysis" is defined broadly, as an all-inclusive activity that could include techniques from a diverse set of domains, including data mining, machine learning, signal and image processing, statistics, and visualization. Consequently, data analysis combines both applied mathematics and computer science with a successful analysis endeavor being a close collaboration among applied mathematicians (including statisticians and machine learning experts), computer scientists, and domain scientists. Such collaboration provides the opportunity for domain scientists to explore available analysis techniques and for the mathematicians and computer scientists to identify problems that need to be solved, possibly involving the extension of current techniques and development of new methods. In addition, a collaborative approach allows the incorporation of domain information into analysis algorithms, leading to more accurate, robust, and faster approaches for solving data-intensive analysis problems.

Many aspects of data analysis must be addressed to meet the requirements of BER, BES, and HEP scientists. While the present focus is on data from experimental and observational facilities, data from computational facilities are included when they are used to support experiments and observations.

Many challenging problems in analysis can be solved better or faster by incorporating domain information. In other problems, such information is essential to solve the problem correctly. Some examples include improving the quality of data, multi-sensor analysis, building models for inference, inverse problems, experiment design, and inference in the presence of uncertainty.

Recommendation: Support a funding model that enables applied mathematicians (including statisticians and machine learning experts) to be embedded with the domain scientists and work closely as an integral part of a team to solve these problems.

4.3.1 Data Quality

Data from experiments and observations frequently have quality issues that can cause problems with the analysis. Often, these data are corrupted by noise from the sensor; convolved by the point spread function of an imaging system; or distorted by extraneous objects, such as clouds in astronomy images. The data can have missing values due to sensors that are inoperable, and there could be spatial and temporal gaps in the coverage resulting from irregular or incomplete sampling. In addition, the data could be contaminated, for example, cell samples in biology. Measurement errors, both systematic and statistical, are invariably present, acquiring increased importance in an era characterized by increases in instrumental sensitivity, as well as area and density of coverage in both space and time.

For some types of data, standard approaches from classical signal and image processing [Bovik05] can be used to reduce the noise, while, in others, there is an opportunity to exploit knowledge of the sensors to devise a domain-specific approach. However, pertinent challenges still remain. For large volumes of data, it can be difficult to select a single algorithm and its associated parameter values

that would be applicable to the data if the noise characteristics vary both spatially and temporally. It often is difficult to ensure that the algorithms do not affect the signal adversely, especially in problems operating close to the sensor's limit. Some noise-reduction algorithms, such as those based on partial differential equations, are computationally expensive. Others, such as image processing techniques, have assumptions of regular sampling, which may not be satisfied by the data. Identification of outliers can be a challenge when it is difficult to define what constitutes an outlier. Finally, as checks for data quality often are the first step in analysis, this initial processing of the data plays an important role in the conclusions drawn from the data and the error estimates associated with the results.

Recommendation: Improving the quality of data prior to analysis typically is the first step in analysis, and it plays an important role in the conclusions drawn from the data and the uncertainty associated with the results. To support this initial data processing, research into algorithms that are robust to accommodate the spatial and temporal variation in the data and designed to account for the characteristics of the sensors are needed.

4.3.2 Improved Statistical, Machine Learning, and Image Analysis Algorithms

Data analysis from BER, BES, and HEP experimental facilities is conducted using a broad spectrum of techniques, ranging from image processing to machine learning and statistics. However, as the complexity of data increases and data sets are explored in new ways, existing techniques often fail to meet science needs. They may be too slow or unsuitable for the data being analyzed or the appropriate algorithms may be lacking. As a result, there is an opportunity for improving algorithms for several tasks, including:

3-D Data Reconstruction: Recreating the 3-D structure of an object from 2-D projections is challenging. First, the low-quality 2-D projections must be identified and removed. This currently is done by a human, an approach that does not scale to large data sets. Second, the reconstruction is obtained by solving an inverse problem, which is not only computationally intensive but poses its own set of challenges (see Section 4.3.8).

2-D and 3-D Image Processing: In many problems, data are in the form of images. These can be 2-D images taken over time with sampling intervals ranging from micro-seconds to days, or they may be 3-D images obtained by taking slices of a 3-D object or through 3-D data reconstruction with a potentially temporal component to the data. The tasks of registration, noise reduction, segmentation, and feature extraction all can be challenging in these problems as the current algorithms may be too slow, inaccurate, or lack robustness to the variation in the data.

Detection of Outliers, Anomalies, and Interesting Events: This task is difficult because there usually is incomplete information regarding what makes an event or object in the data an outlier, anomaly, or interesting event. In some problems, removal of outliers can be done offline to obtain correct statistics on the data, while, in other problems, the detection must be done in real time to enable an alert to be issued. Machine learning and statistical techniques often are used in these tasks, although it is a challenge to reduce false positives and negatives.

Classification and Clustering: Though there is a vast array of techniques for classification and clustering [Bishop07], there still is a need for methods that can handle the data rates of streaming data from LSST or LHC and identify, in real time, events that are interesting and worthy of further observation.

High-dimensional Regression: Techniques from both machine learning and statistics are extensively used in regression problems. Gaussian processes [RW05] are popular when we need to associate an uncertainty with the result, but can be expensive in high-dimensional spaces with a large number of data points. Data compression techniques, such as principal components analysis, can be effective in these circumstances. Alternately, instead of fitting a global model, locally weighted learning [AMS97] can be used to fit a model locally. However, both approaches require the sample points be selected appropriately, especially when the error in the prediction is required to be low. This can be an issue when each sample point, whether the result of an experiment or a simulation, is expensive to generate (see Section 4.3.7).

Streaming Data Analysis: In this class of problems, data are analyzed as they are collected. Typically, the algorithms process the data only in a small window prior to the current time to identify anomalies or interesting events that would prompt an alert. It is a challenge to perform this analysis in real time, especially for multivariate data sampled at different frequencies with possible concept drift while, at the same time, keeping the number of false positives and negatives low.

Improved Sampling: Although the topic of sampling has been around for decades, there still is a need for better sampling approaches that can handle the complexity and high dimensionality of the data. For example, in experiment design (see Section 4.3.7), if sequential sampling is done with a limited number of samples, should all of the samples be run at once, or should there be an alternate between sampling and building surrogate models? As many problems have sample spaces that are high dimensional, how can we ensure that a limited number of samples

span the space adequately? Are there better approaches to sampling than Monte Carlo and Markov chain Monte Carlo (MCMC) techniques [BGJ+11]?

Compression Techniques: While more traditional compression techniques, both lossy and lossless, can provide a solution to handling massive-size data sets, the more recent approach involving compressed sensing [EK12] provides opportunities to optimize the amount of data required to constrain a model or reconstruct a data set.

In many of the preceding algorithms, it is possible to exploit and incorporate domain information to create more accurate and robust algorithms. For example, if it is known that the objects of interest in biological images are round, it may help to detect the objects even in regions of low contrast where segmentation algorithms usually fail. Feature extraction to find suitable representations of objects, such as galaxies or cells, depends on the types of patterns sought in the data. If shape is an important discriminating characteristic, shape features must be considered in the analysis. This not only requires a close collaboration between the domain scientists, applied mathematicians, and computer scientists, but it also indicates a need to develop algorithms that enable the inclusion of domain information.

In analysis, there sometimes is a desire to apply techniques as a “black box,” especially as the data sets grow larger and it is no longer possible to individually examine the analysis results for each item in the data set for correctness. However, correct application of an analysis algorithm requires good understanding of the algorithm, its assumptions, and the correct interpretation of the results. Some techniques may work well on test data but fail on real data, while others sometimes may give unreliable or incorrect results. As new algorithms are developed to address analysis requirements for BER, BES, and HEP science-use cases, there is a

need for algorithms that are both interpretable and robust to any spatiotemporal variation in the data. A close collaboration between ASCR and domain scientists would further ensure the careful and considered application of analysis techniques.

Recommendation: As the volume, velocity, and variety of data from experimental facilities increase and the data collected are uncertain and imprecise, classical techniques for analysis developed by the statistics, machine learning, and image processing communities are no longer sufficient to address scientific analysis needs. Advances in algorithm research are required to address these gaps, coupled with increased collaborations between ASCR and domain scientists to ensure the relevance of this research.

4.3.3 Approximate and Automated Algorithms

The increasing size of data from computational, experimental, and observational facilities has resulted in new requirements for algorithms. In some problems, such as nearest neighbor searches or sequence alignment, the task's computational complexity is such that it becomes computationally infeasible for large, high-dimensional data sets. In these cases, approximate algorithms are of interest. These algorithms return an approximate result in a short time. As the algorithm is allowed to run longer, more accurate results are obtained. This tradeoff between accuracy and speed introduces errors in the analysis process, so these must be quantifiable for the algorithms to be used in practice [AMN+98].

When the large size of a data set is coupled with the variability across the data set, a different requirement for analysis algorithms arises. Ideally, researchers would like to select the algorithms and parameters for processing the data and apply them to the entire data set. If the variability across the data set is large, this

may yield incorrect analysis results. For example, when many thousands of images are analyzed, each with varying quality, a fixed algorithm with a defined set of parameters is likely to be unsuitable for analysis. Therefore, automated approaches that can adapt to variation in the data are desired [RBU08].

Recommendation: The increasing size and complexity of data imply that some computationally expensive algorithms will no longer be viable, and a single algorithm is unlikely to work given the data variability. Approximate versions of algorithms, as well as adaptive selection of algorithms and parameters based on the data, are required to fill this gap.

4.3.4 Multi-sensor and Multi-resolution Analysis

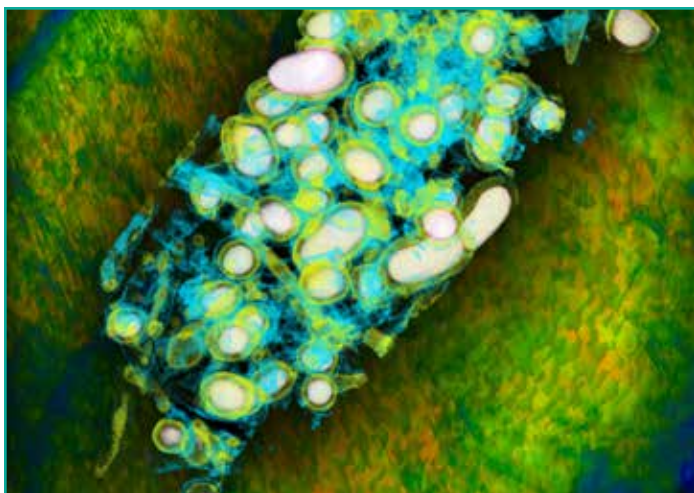
In many application areas, scientists perform different experiments or computer simulations that are linked across time or length scales or provide different physics insights into a phenomenon. For example, in designing materials for fuel cells, scientists at the APS are interested in combining nanometer-resolution X-ray tomography images with micrometer-resolution movies of the fuel cells at work to investigate phenomena such as crack propagation. Similar examples arise in cosmology, where observations obtained at varying frequencies might be combined, and in the analysis of biological or genome data from multiple measurements, where there is interest in studying how RNA in samples changes over time. Finally, some experiments using focused light spectroscopy produce data sets that span multiple spectra. Herein, analysis involving multiple length or time scales or multiple physics is referred to as “multi-modal” analysis. Although many approaches have been suggested in the data fusion [LHL08] field, there is a need for computational techniques tailored to the types and volumes of data collected by BER, BES, and HEP scientists to allow them to extract knowledge from a set of diverse experiments.

The fusion of multi-modal data calls for the development of new mathematical models that combine data and simulation from multiple scales and provides an opportunity to include physical models in data analysis. Often, the data-to-data and data-to-model matching problem can be formulated naturally as constrained optimization problems, where the constraints capture scientific phenomena, expert knowledge, or data and model inaccuracies. Research is required into scalable, large-scale, constrained optimization techniques that also can take the inherent uncertainties into account, either through stochastic programming [BL97] or robust optimization techniques [BTN02]. Other relevant techniques include: multiscale modeling, image registration [Modersitzki04], time-series analysis, Bayesian approaches, and Dempster-Shafer theory [LHL08]. Advances in these areas are necessary to fully exploit the science in large collections of heterogeneous data sets.

Recommendation: Promote multidisciplinary research into techniques for exploiting multi-modal data, such as data fusion methods and multiscale constrained optimization. These techniques would incorporate physical insight, quantify uncertainties, and run efficiently on emerging architectures.

4.3.5 Visualization

The scale and complexity of today's computational, experimental, and observational facilities already overwhelm existing visualization tools. As the volume of data continues to increase, some data exploration tasks, presently performed manually, become too time-consuming and tedious for scientists to carry out. To eliminate as many manual tasks as possible, it is imperative to develop intelligent data analysis methods—coupled with interactive visualization—through easy-to-use user interfaces. Existing tools lack both interactive performance and intuitive interfaces that are essential for tasks that explore the



Using soft X-ray microtomography, this image depicts high-resolution, reconstructed biofilm cells.
(Courtesy: PNNL)

data, isolate and verify features of interest in the data, and conduct comparative visualization and analysis. To support this, new algorithms and acceleration techniques are required. User interfaces and visual representations should be customized to suit science workflows and match domain languages. The resulting tools should be evaluated with usability studies and thoroughly tested for their robustness.

Most large-scale scientific investigations, such as the LSST and LHC, are highly interdisciplinary and collaborative with project investigators often geographically distributed. There is clear benefit for these scientists to conveniently share their data, knowledge, and research findings without time/place constraints, which emphasizes the need for remote access to support such collaborations. Furthermore, sharing should extend beyond browsing the stored information. A collaborative visualization system [IES+11] that can automatically extract associations in the data and make recommendations to the scientists is needed. Again, to enable such a system, the design of an easy-to-use remote visualization interface is crucial. The interface design should account for different visual means to present the requested information, desired operations, and even social aspects of the collaboration.

In the case of light sources, photon sources, and nanoscience facilities, real-time online data processing and efficient offline data analysis are challenging tasks due to the algorithmic complexity. However, to achieve informed decisions and fully exploit the obtained beam time, it is vital for the team conducting the experiment to have the necessary real-time data visualization. The increasingly fast scanning functionality of modern devices requires equivalent efficient online data processing. For example, X-ray fluorescence microscopy can generate thousands of 2-D projection images (50 gigabytes/per second) that need to be processed in real-time to afford 3-D reconstruction.

Historically, visualization has focused on creating images or movies of data. With the growth in data size and complexity, it is increasingly the case for both experimental and computational sciences that visual and numerical quantitative analyses offer traction on the data-intensive problem. An expected trend toward a greater coupling, or linking, between quantitative methods, such as statistical analysis, feature detection/tracking/analysis, and visual data exploration and analysis techniques, will prove invaluable in scientific knowledge discovery.

Over the years, visualization and analysis tools have focused on techniques that are applicable to a single data set. Some areas of science, especially climate modeling, routinely produce ensemble collections, where the focus is to study the characteristics and trends across multiple data sets. The trajectory of experimental sciences is toward similar studies involving properties of data collections. To that end, opportunities exist for enabling scientific insight through a combination of new technology development along with application of promising, existing techniques in the areas of ensemble analysis, uncertainty quantification and analysis, statistical analysis, and visualization. The scope of work here

potentially is broad and could include analysis and visualization techniques and methods in graph-based processing, multivariate spatiotemporal and multiscale processing, and comparative studies.

Recommendation: Research and development of visualization methods and tools for observation and experimental data are required. This includes development of methods for ensembles, multivariate, and multiscale data. Such quantitative methods should support collaboration and be scalable and interactive.

4.3.6 Scalable Parallel Algorithms

The analysis of data generated by BER, BES, and HEP computational, experimental, and observational facilities will require parallel implementations of many algorithms. One of the main motivating factors is the large size of the data. While some analysis tasks can be embarrassingly parallel, such as the processing of images taken of different parts of the sky, others, such as the widely used pairwise sequence alignment algorithms in genomics, are not readily parallelizable. Another motivating factor is the need for real-time response in some tasks, such as alerts in astronomy surveys, or the need to modify an experiment on the fly based on the outcome of the previous experiment/measurement, which must be analyzed in limited time. A third factor is the geographically distributed nature of the data in domains such as climate, where the analysis algorithms must be modified to analyze the distributed data and return the same results as if the data were collocated. In addition to parallel and distributed algorithms, some domains, such as astronomy, also require scalable data structures, as well as a better understanding of which data structures are suitable for which algorithms, so the appropriate data structure selections can be made for improved performance. If data structures that

provide optimal performance across a range of analysis algorithms can be identified, the optimization of common operations can be done only once, providing solutions to several problems.

These scalable algorithms and data structures should be implemented to fully utilize the hardware power provided by emerging high-performance computers [ABD+09]. In particular, parallelization of existing and new image algorithms on general-purpose graphics processing units (GPGPUs) and hybrid multi-/many-core co-processing will dramatically decrease processing and analysis time and improve facilities' efficiency. The moderately parallel machines likely to be collocated and used by the experimental facilities are expected to have architectures similar to extreme-scale systems [KBB+08]. Thus, they will be equally challenging to program. These multi-core machines usually are characterized by highly heterogeneous organizations with non-uniform memory access (NUMA) nodes, enhanced with GPU-type accelerators and/or single instruction multiple data (SIMD) fine-grained parallelism. They are expected to have smaller memories, relatively lower bandwidths, and a high cost for data movement. For example, various image processing techniques often used in X-ray data processing and visualization, such as peak localization and fitting for microscopy, 3-D tomography reconstruction, and differential phase contrast microscopy, can benefit directly from advanced parallel software that takes advantage of co-processing/GPGPUs. Similarly, image processing algorithms in other domains, such as cosmology observation and transmission electron microscope images, can take advantage of these hybrid architectures. The optimal and scalable implementations of analysis algorithms will require optimizing data locality and supporting re-use. Reducing data movement, especially for data-intensive problems, will be a challenge.

Recommendation: Research in parallel analysis algorithms and data structures that are scalable should be conducted, especially on emerging high-performance architectures, to continue to meet analysis needs of data-intensive problems that require a fast turnaround or work with geographically distributed data.

4.3.7 Experimental Design

The experimental design problems that arise in data-intensive science can be formulated as “resource allocation” problems, e.g., given a limited budget (time, computation, experimentation), where and how should additional information be collected? This may be a choice of which angles/projections to record in a tomography problem, which experiments to run to improve model-based predictions, or which input settings to use for simulation model runs to best make a prediction about climate. Typically, the choice about where to collect additional information is made to maximally reduce uncertainty in a quantity of interest. This, of course, relies on methods to understand the current state of uncertainty about the system (refer to Section 4.3.9). In many settings, approximate criteria can be used to afford useful designs without undue amounts of computation, such as space-filling designs [CL07].

When multiple sources of information are available for the analysis, determining good or optimal designs requires some method of trading off different information sources. This is the case when considering multi-modal data (see Section 4.3.4). For data-intensive science, the “standard” design of an experiment problem can expand to encompass choosing among the entire set of multi-modal data collection opportunities.

Just as model output is reduced and aggregated (e.g., seasonal averages over a 100-km² grid), measurement data also are aggregated, reduced, and summarized, such as:

- Aggregations over space and time, matching model output
- Real-time searching for “trigger” events to begin more thorough data collection
- Processing of raw sensor measurements into inferred physical measurements (e.g., turning spectral reflectance measurements from a satellite into CO₂ concentrations or turning camera pixel measurements into particle tracks at the LHC)
- Detecting anomalies, or important events, in streaming data.

Measurement data reduction/aggregation is necessary in inference because it makes the computations required for parameter estimation, sensitivity analysis, and prediction feasible. The question of how the data should be aggregated/reduced is influenced by considerations regarding hardware, computational modeling, statistical analysis, and the application itself. While no general theory or framework exists for aggregating/reducing measurement and model output, examples in fields such as climate, particle physics, and cosmology offer a variety of successful approaches. The question of determining optimal aggregation/reduction schemes could be integrated into the experiment’s design, seeking schemes that either optimize uncertainty reduction or, in some problems, optimize the tradeoff between uncertainty reduction and the robustness of the results obtained.

Recommendation: In large-scale, data-intensive scientific investigations, experimental design considerations may range from finding the best input settings for a computational model run to optimizing the end-to-end scientific investigation, considering hardware, data collection, and analysis choices. Advances for these design methodologies are needed to tackle more complicated applications of resource allocation and experimental design while accounting for cost, accuracy, resilience, and other decision metrics.

4.3.8 Inverse Problems

Many data sets collected from BER, BES, and HEP computations, experiments, and observations are used to deduce the structure and/or dynamic properties of physical systems that cannot be directly measured. The solution of this inverse problem often is formulated using optimization, where minimizing the discrepancy between a forward model and the measured data with respect to a set of decision variables is required. However, to be physically meaningful, the solution generally involves the determination and exploration of the full posterior distribution of a probability distribution over a set of variables rather than simply the location of the peak.

Inverse problems usually are ill-posed for a variety of reasons, making them difficult to solve even when a large volume of measurement data is available. The measurements may be indirect and/or incomplete, their quality may be variable, and the measurement errors and likelihood function may not be well-determined or of a simple form (e.g., Gaussian). Regularization techniques used to overcome the ill-posedness must be chosen carefully with knowledge of the forward problem.

To effectively solve inverse problems, a sufficiently complete forward model must be established that offers accurate descriptions of the measurements by accounting for

the processes that describe the measured phenomena, details, and uncertainties in the experimental setup, instrument errors, and systematic and stochastic nature of noise and other sources of contamination (e.g., foregrounds and backgrounds). In addition, because many runs of the forward model may be required (e.g., 10,000–100,000 in the case of MCMC), the solution must be made computationally tractable by improving the computational efficiency and reducing the number of model evaluations without losing the forward model's fidelity.

There has been significant progress in solving linear inverse problems using methods based on total variation minimization [ROF92] and iterative least squares solvers [Hansen97], combined with appropriate regularization techniques. In some problems, non-uniform fast Fourier transforms have improved solution accuracy, while compressive sensing techniques are enabling the solution to problems that have sparse representation, even when these problems first may appear to be underdetermined. However, inverse problems still can be challenging when the data volume is large and data quality is variable and/or poor. Acceleration and preconditioning techniques are needed to reduce the solution time, while image processing techniques (e.g., clustering and filtering) can be exploited for more robust solutions.

Additional challenges arise when solving nonlinear inverse problems. These include the single-molecule diffractive imaging problem [NWV+00, MCK+99, SD04, CBM+06], phase retrieval [MHC+03], single-particle cryo-electron microscopy problem [Frank06], and calibrating cosmological parameters [HHH+07]. Maximum likelihood formulations of restricted versions of these problems can be solved using iterative methods, although the convergence of these methods can be slow when the Hessians have slowly decaying singular values. However, these typically non-convex problems require a good

starting guess to avoid being trapped at a local optimum. More general approaches involve estimating and exploring the full posterior distribution (usually via MCMC, although other MC samplers can be used), especially when accurate and robust error estimation is crucial. In some, more recent data-intensive applications, the statistical quality of the data is such that errors from the methods themselves can be more significant than the measurement errors.

To reduce the total number of nonlinear iterations, new fast algorithms also are needed, for example, using alternative formulations, such as the alternating direction methods developed in [WYL+12] for solving phase retrieval problems, or through convex relaxation techniques to address convergence issues [CES+13]. Good initial guesses may be obtained by using dimension reduction and machine learning to detect the problem's underlying structure.

Finally, Bayesian methods provide an alternate approach to solving inverse problems by using the prior information for regularization. Given the strong dependence of any solution approach on the characteristics of the underlying inverse problem, solving these problems will require a close collaboration among applied mathematicians, statisticians, and domain scientists.

Recommendation: Inverse problems are a particularly challenging class of analysis problems given their ill-posedness, the computational cost of any solution, and the need to exploit domain information to make the problem tractable. A multidisciplinary approach is recommended, leveraging expertise from both ASCR and domain scientists.

4.3.9 Inference, Prediction, and Reasoning under Uncertainty

Modern data collection and computational modeling have enormous potential to advance understanding in a variety of complex systems—

physical, biological, or social. This understanding will be advanced by thoughtful combination of the following:

- Vast amounts of data, both from physical observations and computational modeling
- Insight and reasoning from application science
- Methodology from statistics, machine learning, applied math, and other related fields.

Appropriate accounting for uncertainty and error from various sources is critical to making useful inferences. These include experimental and measurement error, sampling variability, the choice of the theoretical model used to predict the quantity of interest, uncertainty in model inputs, the adequacy of the theoretical model, and approximations arising from the computational implementation of a given model. It is necessary to characterize this uncertainty in an appropriate form within the data representations to enable queries on the uncertainties and propagate the uncertainties via the analysis.

In data-intensive science, the “standard” issues regarding inference, prediction, and reasoning under uncertainty are expanded. Considerations regarding data reduction (Section 4.1.2), data storage (Section 4.2.5) and movement (Section 4.2.1), and computational constraints, also must be considered in developing inferential methodology and algorithms (Section 4.3.2). Here, interactive tools, such as visualization (Section 4.3.5) and other exploratory data analysis capabilities, are crucial for developing and assessing appropriate data reductions, as well as potential algorithms for extracting information from vast amounts of data.

Many analyses will need to consider huge amounts of model output, as well as substantial amounts of output from sensors, detectors, and/

or other measurement processes. Data allow inference about key features of the physical processes of interest:

- Did we see the Higgs boson particle?
- How fast is the universe expanding?
- How will a particular material respond in an extreme radiation environment?

As such, data typically help reduce uncertainty regarding predictions and key model parameters. Principles, including likelihood and sampling, commonly are used to link the data to the inferences being sought. However, as large, complex data are aggregated/reduced, applying such principles becomes more difficult. In some cases, the observational process can be modeled, from physical emissions to the eventual sensor signal. In these cases, ideas from approximate Bayesian computation [SF12] could be adapted to link the measurement to the physical process of interest. Regardless, there is a variety of needs and opportunities in relating measurement data, perhaps with substantial reduction or under streaming conditions, to the inferences desired for the investigation at hand.

In addition, instead of a single experiment, multiple modalities and representations of data may need to be accommodated. We are considering not just experimental and measurement errors that arise from raw data, but how those errors are propagated as the data are reduced, features are extracted, and analysis is performed. Similarly, we are not considering the “standard” issue of adequacy for a single model but, instead, are analyzing a series of models, from the “best” (highest fidelity) to a reduced model. Linking these considerations to specific scientific problems allows concrete exploration, accounting for physical constraints, modeling choices, and computational and data limitations.

Recommendation: The task of connecting data to scientific models for inference in emerging data-intensive environments will need to balance mathematical and statistical considerations with those of computational speed and veracity, as well as data storage, movement, and velocity. Research and examples in this area will help identify fruitful directions and promising pathways for data-intensive inference.

5 Global Themes

In this section, issues that are of a general nature and cut across all areas are discussed. They are referred to as “global themes.”

5.1 Cost-model-based Data Processing System Design and Operation Optimization

Data processing systems are becoming increasingly complex, incorporating a variety of heterogeneous computer systems architectures, operating systems, programming models, and software solutions. With increasing data volumes, it becomes essential to optimize system performance in terms of speed, throughput, and reliability. Unfortunately, such an optimized design might not be affordable within the resources of a given user facility. Thus, it is equally important to optimize the design and its long-term operation in terms of resources required for hardware, software, energy, and staff costs. Facility staff must be able to assess different solutions and weigh the necessary tradeoffs to make decisions with high financial and scientific impact.

Recommendation: Develop “cost models” that allow users to evaluate, compare, and optimize designs, both for specific processes and end-to-end applications, as well as incorporate hardware (instrument and computing), software, data, and networking.

5.2 Human Computer Interface

Over the past decades, the ASCR community has developed many outstanding tools—often incorporating unique optimizations for data processing and analysis at the extreme scale. Unfortunately, these tools usually are not easy to use or adapt to the needs of less-experienced

experimental scientists who visit user facilities for a short period of time and need to be productive straightaway as access time to large-scale facilities is limited. Little emphasis has been placed on studying the fundamental factors that may influence usability and adoption of software tools in terms of enabling scientific productivity and support of insight generation and discovery.

Recommendation: More emphasis should be placed on the human computer interface (HCI) component of software development and delivery.

Recommendation: Require DOE software development proposals to address explicitly the question of how software will be adopted by user communities.

5.3 Software Quality, Resilience, and Readiness

Any software developed for data processing systems must be deployed in an operational environment, running potentially 24/7 for years. This requirement results in stringent needs for quality, resilience, and readiness. A key challenge in the development of sustainable software solutions for production use is the transition from research product to production-ready tool. Currently, no clear organizational or funding path to accomplish this process exists.

In the business world, software-as-a-service (SaaS) methods increasingly are used to reduce costs and improve robustness, resilience, and the capability of software delivered to remote users. Similar approaches have been applied successfully in the scientific world via projects such as MG-RAST⁴¹, KBase, nanoHUB⁴², and Globus Online⁴³. There appears to be opportunities to apply such methods far more extensively in science.

⁴¹ <http://metagenomics.anl.gov/>

⁴² <http://nanohub.org/>

⁴³ <https://www.globusonline.org/>

Recommendation: Investigate the applicability of ASCR exascale resilience work for the design of fault-tolerant, resilient data processing solutions.

Recommendation: Apply verification, validation, and uncertainty quantification methods in the development of data processing solutions.

Recommendation: DOE programs jointly need to devise a clear funding and responsibility path from research to production-level software that includes the long-term sustainability of operational software.

Recommendation: Support both research investigations and production deployments of SaaS solutions to the challenges of software quality, resilience, and robustness.

5.4 Enhance Use of Modeling and Simulation for Experimental Design

In several of the breakout sessions, there were discussions noting the importance of HPC modeling and simulation in data-intensive science across the Office of Science programs. Modeling and simulation already are used to design instruments and better understand experimental results. However, the point was made that more extensive use of modeling and simulation in the context of future project planning would enable science teams to better understand system characteristics and costs and evaluate alternatives. This discussion spawned two recommendations:

Recommendation: Research and development should be pursued to improve the availability of modeling and simulation tools for use in experimental design for data-intensive science projects, ranging from facility-level architectures to wide-area, end-to-end scientific workflows.


Recommendation: Data experts in the Office of Science should meet regularly to exchange ideas and experience related to the many aspects of data-intensive science.

5.5 Sharing Modeling, Simulation, and Analysis Tools

Another major discussion theme examined how the ASCR community can support work among other offices. Some approaches identified include research, development, and consulting support. Participants requested a centralized location for engaging interested collaborators in solving BES, BER, and HEP issues. Requests were made for custom interfaces and tutorials on ASCR solutions to improve their usability for specific scientific communities. How to improve communication by learning each other's "scientific language" also was discussed. To increase broad usability, it was suggested that ASCR focus its effort on web-based interfaces and on streamlining security procedures.

One interesting idea that resonated with the workshop group was a collection of experimental data challenges as an integral part of ASCR facility data test beds. These data challenges might take the form of an innovation competition, such as the Netflix Prize competition. Such public, crowd-sourced competitions are encouraged by the Office of Science and Technology Policy (OSTP) to engage the broader community in data-intensive problems of interest to the nation.

Recommendation: ASCR should immediately engage facilities, test beds, outreach, and research and development efforts to focus on supporting the Office of Science's experimental data community. The workshop group also encouraged rewarding customization, usability, and reuse of ASCR solutions by other offices and recommended supporting the creation of publicly available data challenges and innovation competitions that characterize and provide frameworks for solutions to solve DOE experimental data problems.



Recommendation: Research and development should be pursued to improve the availability of modeling and simulation tools used in experimental design for data-intensive science projects, ranging from facility-level architectures to wide-area, end-to-end scientific workflows.

Recommendation: Data experts from the various DOE laboratories in the Office of Science should meet regularly to exchange ideas and experience related to the many aspects of data-intensive science. This exchange can take the form of a yearly workshop followed by a series of monthly conference calls to formulate the details regarding recommendations and progress.



6 Summary of Crosscutting Findings and Recommendations

6.1 Findings

1 FINDING 1:
The challenges associated with scientific data are diverse and often distinct from challenges in other data-intensive domains, such as web analytics and business intelligence.

The volume and velocity of scientific data can be extremely high. Scientific data are precious and can be impossible or expensive to regenerate. Transparency and access to scientific data are important considerations. Tools and technologies developed for other applications likely will be insufficient to address all of the data science needs required by the Office of Science.

2 FINDING 2:
Research communities across the Office of Science have considerable expertise in the aspects of data science necessary for performing their science.

For example, HEP is excellent at real-time data ingestion, analysis, and distribution due to the needs of accelerator and astrophysics facilities. The climate community in BER and its ASCR partners are world-class at data curation, provenance, and access-control because of the close scrutiny their science receives. BES facilities excel at local data reduction techniques. Meanwhile, the ASCR community is outstanding at data analysis and visualization due to its experience with large simulation data sets.

However, the data science communities in different parts of the Office of Science are not fully aware of each other's capabilities and often do not coordinate their activities. This

can lead to inefficiencies and missed scientific opportunities. Greater coordination and communication across the Office of Science—in headquarters and among researchers—would be beneficial.

3 FINDING 3:
Many Office of Science experimental facilities anticipate rapid growth in data volume, velocity, and complexity.

The Office of Science experimental facilities representatives expressed considerable concern that existing technologies will be insufficient to address upcoming data challenges. They need end-to-end systems that provide more automated workflows and capabilities to ingest, analyze, and manage much larger and more complex data sets generated at faster rates. Many of the core needs are similar across different science facilities, but the detailed requirements can be specific to each facility.

The rapid growth of data rates will require advances in analysis techniques to address real-time decision making, near-real-time data reduction, and the challenge of analyzing larger data sets offline. These needs will require advances in statistics, machine learning, visualization, and related areas. Progress will require that mathematicians and computer science researchers work closely with domain scientists.

4 FINDING 4:
Currently, many scientific facilities expect users to manage their own data.

This is particularly true of facilities that support a large number of diverse experiments, e.g., light

sources, nanoscience facilities, and neutron sources. A greater degree of centralized support for data management, analysis, storage, and remote access would have a number of advantages. It would help address the challenges of impending data growth, enhance efficiency by reducing duplication of effort, and provide more consistent analysis and higher-quality archival support that would create new scientific opportunities, as well as provide a mechanism for open access.

5 FINDING 5: **There is an urgent need for standards and community APIs for storing, annotating, and accessing scientific data.**

The development of standards and protocols for distributed data and service interoperability is essential. Furthermore, API standards will enable collaborations and facilitate extensibility, whereby similar, customized services can be developed across science domains. Such standardization will facilitate data reuse and integration from multiple experiments. It also will be needed as part of any move to provide facility-wide data services.

6.2 Recommendations

1 RECOMMENDATION 1: **The Office of Science should support multidisciplinary teams to conduct research and development needed to address DOE's unique data science challenges.**

The data challenges confronting the Office of Science facilities can only be addressed by the combined efforts of computer scientists and mathematicians working closely with facilities experts and domain scientists. To have substantive impact, research projects must be

deeply informed by the complex needs of DOE's experimental sciences. The following areas are high priorities for investment:

- Flexible infrastructure for data management, curation, storage, and remote access that can be shared across communities
- Efficient methods for data reduction, storage, and access
- Scalable methods for data analysis, including statistics, machine learning, and visualization
- Techniques for combining data from multiple experiments
- Modeling capabilities to support the optimal design of data management systems
- Techniques for using simulations in support of experiments and applying experimental data to validate simulations
- Services that allow for low-cost, intuitive access to powerful data collecting, management, analysis, curation, and sharing capabilities.

2 RECOMMENDATION 2: **DOE science facilities should provide more centralized support for data management, storage, analysis, and access.**

A number of DOE science communities use facilities as throughput resources. Many small science teams generate tremendously diverse scientific results. Often, these teams are expected to manage their data with limited assistance from the facilities. In the future, this approach likely will be untenable. Facility enhancements will dramatically increase data volumes. Greater emphasis on scientific transparency will require open access to data, and new science will undoubtedly be discovered by making connections across and between experimental data sets. All of these

drivers underscore the need for facilities that provide policies and technologies offering greater centralized support for data challenges.

3 RECOMMENDATION 3: **The Office of Science should develop a cross-organizational strategic plan for data science.**

Data science cuts across communities and is broader than any single component with the Office of Science. As such, coordination is essential. The plan should provide a framework for investment and prioritization with each office and identify dependencies between them. The topics that should be addressed include: data sharing policies, data curation standards, data science facilities and services, and sustainable software development and deployment. Such a plan would lead both to improved efficiencies and scientific productivity.

4 RECOMMENDATION 4: **Mechanisms should be created to enhance communication among the scattered data science communities within the Office of Science.**

There will be significant benefits from exchanges of experience, best practices, perspectives, and current challenges.



7 Glossary

7.1 Terms

Classification, regression. A class of techniques, which, starting with a set of data items, each described by several features or characteristics and an associated output, builds a predictive model that can assign an output to an unseen data item, given its features. A discrete (continuous) output corresponds to classification (regression) technique.

Clustering. A class of techniques, which, starting with a set of data items, each described by several features or characteristics, groups the items into clusters so items in a cluster are more similar to each other than to items in a different cluster.

Collaborative visualization. The shared use of computer-supported (interactive), visual representations of data by more than one person with the common goal of contribution to joint information processing activities.

Comparative visualize. Techniques to visually compare similarities and differences between data sets.

Concept drift. The phenomena where statistical properties of the data change over time while remaining within normal conditions.

Data Analysis. Techniques and tools to extract knowledge from the data. These include: methods and algorithms for enhancing data quality, various statistical and machine learning techniques, multi-resolution and multi-sensor analysis methods, and large-scale visualization techniques.


Data Management. Activities that tend to storing, searching, and sharing data. These include: I/O acceleration to storage systems, data retention techniques, tools for data sharing for the communities involved, search tools for identifying subsets of interest, and tools that support data models for representing the domain view of the data.

Data Processing. Activities that must take place while data is collected from experiments/ observations. These include challenges with data acquisition, data reduction, data transformation for subsequent analysis, data movement to remote sites (where data is stored), workflows for multiple tasks pipelines, and automatic collection of metadata and provenance about the data being collected.

Experimental design or design of experiments. The design of any information-gathering exercise where variation and/or uncertainty are present. For physical experiments, there are three broad principles for experimental design: randomization, replication, and blocking. For deterministic computer codes, these principles do not apply, and the design points are chosen to maximize some other criteria, for example, to explore as much of the input region as possible or to give the best-expected prediction accuracy.

Interactive visualization. A real-time process of transforming and viewing scientific data as visual representation.

Reduced model. A lower-fidelity model developed to replace (or augment) a computationally demanding, high-fidelity model. A reduced model is sometimes called an emulator, although an emulator sometimes refers specifically to the use of a response surface as a reduced model.



Response surface. A function that predicts outputs from a model as a function of the model inputs. A response surface typically is estimated from an ensemble of model runs using a regression, Gaussian process modeling, or some other estimation or interpolation procedure.

Uncertainty quantification. The process of quantifying uncertainties in a computed quantity of interest with the goal of accounting for all sources of uncertainty and quantifying the contributions of specific sources to the overall uncertainty. More broadly, uncertainty quantification can be thought of as the field of research that uses and develops theory, methodology, and approaches for carrying out inference, with the aid of computational models, on complex systems.


Usability study. Techniques to evaluate a software tool's quality by measuring user responses as they use the tool to complete a series of tasks.

7.2 Acronyms

ACL	Access Control List
ADARA	Accelerating Data Acquisition, Reduction and Analysis (Lab supported project at ORNL. ADARA develops a streaming data workflow between the SNS facility and OLCF.)
ADIOS	Adaptive IO System (A tool that provides a common layer for I/O services at runtime.)
ALS	Advanced Light Source (BES user facility. ALS is a synchrotron located at LBNL.)
API	Application Programming Interface
APS	Advanced Photon Source (BES user facility. APS is a synchrotron located at ANL.)
AR5	Assessment Report, Fifth (Fifth IPCC Assessment Report, with publication planned in late 2013.)
ARM	Atmospheric Radiation Measurement (ARM) Climate Research Facility (BER user facility for climate research. ARM is located at multiple laboratories.)
ATLAS	A Torroidal LHC Apparatus (One of four particle detectors for the Large Hadron Collider at CERN.)
BELLE	An Intensity Frontier experiment at KEK, Japan. PNNL is leading the U.S. contribution to the Belle II upgrade.
BLAST	Basic Local Alignment Search Tool
CFN	Center for Functional Nanomaterials (One of the five BES NSRC user facilities. CFN is located at BNL.)
CINT	Center for Integrated Nanotechnologies (One of the five BES NSRC user facilities. CINT is located at SNL (NM) and LANL.)
CMB	Cosmic Microwave Background
CMS	Compact Muon Solenoid (One of four particle detectors at the Large Hadron Collider.)
CNM	Center for Nanoscale Materials (One of five BES NSRC user facilities. CNM is located at ANL.)
CNMS	Center for Nanophase Materials Science (One of five BES NSRC user facilities. CNMS is located at ORNL.)
EBMC(s)	Electron Beam Microcharacterization Center(s) (A collection of three BES electron beam characterization user facilities.)
EMC	Electron Microscopy Center (One of three BES EBMCs. EMC is located at ANL.)

EMSL	Environmental Molecular Sciences Laboratory (A BER user facility for the environmental and molecular sciences. EMSL is located at PNNL.)
EOS	Earth Observing System (A NASA project of coordinated series of polar-orbiting and low inclination satellites.)
ESnet	Energy Sciences Network (ASCR user facility. ESnet provides a high-bandwidth network for the national laboratories, universities and other research institutions.)
GEANT4	Geometry And Tracking 4 (A toolkit for the simulation of the passage of particles through matter.)
GPGPU	General purpose GPU
HCI	Human Computer Interface
HDF5	Hierarchical Data Format 5
HFIR	High Flux Isotope Reactor (A BES user facility. HFIR is a flux reactor based neutron source located at ORNL.)
HPSS	High Performance Storage System (an archival mass storage system.)
HYSPEC	Hybrid Spectrometer (An SNS beam line that combines SNS time-of-flight technique with a crystal spectrometer for neutrons.)
IPCC	International Panel on Climate Change
JGI	Joint Genome Institute (A BER user facility providing genome sequencing, acquisitions, and analysis. JGI is located in Walnut Creek, California.)
KBase	Systems Biology Knowledgebase (BER supported software and data environment for systems biology.)
LCLS	Linac Coherent Light Source. (BES user facility. LCLS is a synchrotron located at SLAC National Accelerator Laboratory.)
LHC	Large Hadron Collider (LHC is a particle accelerator located at CERN near Geneva, Switzerland.)
LSST	Large Synoptic Survey Telescope (An HEP supported telescope planned to be located in Chile.)
MCMC	Markov Chain Monte Carlo
MC	Monte Carlo

MeDICI	Middleware for Data-Intensive Computing (Platform for developing distributive streaming analytics. MeDICI was developed at PNNL.)
NCCR	National Center for Computational Science (Scientific computing center at ORNL.)
NCEM	National Center for Electron Microscopy (One of the three BES EBMCs. NCEM is located at LBNL.)
NERSC	National Energy Research Scientific Computing Center (ASCR HPC user facility. NERSC is located at LBNL.)
NetCDF	Network Common Data Format
NEWT	Nice and Easy Web API for HPC (A NERSC collection of applications that allow scientists and the public to write web apps for HPC.)
NeXus	Neutron X-ray and Muon Science (A developed international standard for a common data format for x-ray, neutron, and muon science.)
NoSQL	No SQL (Database systems that do not comply with the SQL relational databases standard and provide flexible and simple data model.)
NSLS	National Synchrotron Light Source (A BES synchrotron user facility. NSLS is located at BNL.)
NSLS-II	National Light Source II (A BES synchrotron user facility. NSLS-II is the next generation NSLS and replaces NSLS at BNL.)
NSRC	Nanoscale Science Research Center (A collection of five BES user facilities for nanoscale science research.)
NUMA	Non-Uniform Memory Access
ROOT	An object oriented framework for data processing and analysis developed by CERN.
Science DMZ	Science Demilitarized Zone (A development of ESnet to optimize data movement across network perimeters of data transfer end-points.)
ShaRE	Shared Research Equipment (One of the three BES EBMCs. ShaRE is located at ORNL.)
SIMD	Single Instruction, Multiple Data
SMRT	Single Molecule Real Time sequencing (A DNA sequencing methodology.)
SNS	Spallation Neutron Source (One of two BES Neutron Facilities.)
SQL	Structured Query Language



SSRL	Stanford Synchrotron Radiation Light Source (A BES synchrotron user facility. SSRL is located at SLAC National Accelerator Laboratory.)
Swift	A Scalable Workflow language and system developed at ANL and U. Chicago
TMF	The Molecular Foundry (One of the five NSCRs; BES user facility at Lawrence Berkeley National Laboratory.)
uclust	Ultrafast Cluster program
VO	Virtual Organization (Organization of physically distributed collaborations.)
WAN	Wide Area Network
WCRP	World Climate Research Program
WGCM	Working Group on Coupled Modeling

8 References

- [ABC+12] B. Allen, J. Bresnahan, L. Childers, I. Foster, G. Kandaswamy, R. Kettimuthu, J. Kordas, M. Link, S. Martin, K. Pickett and S. Tuecke. "Software as a Service for Data Scientists." *Communications of the ACM*, 55(2):81-88, 2012.
- [ABD+09] K. Asanovic, R. Bodík, J. Demmel, T. Keaveny, K. Keutzer, J. Kubiatowicz, N. Morgan, D.A. Patterson, K. Sen, J. Wawrzynek, D. Wessel and K.A. Yelick. "A view of the parallel computing landscape." *Communications of the ACM*, 52(10):56-67, 2009.
- [ADH+12] D. Asner, E. Dart, and T. Hara (eds.). Belle-II Experiment Network Requirements Workshop. Final Technical Report, LBNL-6268E, 2012. Available online at: http://www.es.net/assets/pubs_presos/Belle-II-Experiment-Network-Requirements-Workshop-v18-final.pdf.
- [AMN+98] S. Arya, D.M. Mount, N.S. Netanyahu, R. Silverman and A.Y. Wu. "An optimal algorithm for approximate nearest neighbor searching fixed dimensions." *Journal of the ACM (JACM)*, 45(6):891-923, 1998.
- [AMS97] C.G. Atkeson, A.W. Moore and S. Schaal. "Locally Weighted Learning." *Artificial Intelligence Review*, 11:11-73, 1997.
- [BGJ+11] S. Brooks, A. Gelman, G.L. Jones, and X-L Meng (eds.). *Handbook of Markov Chain Monte Carlo*, Chapman & Hall/CRC, 2011.
- [Bishop07] C.M. Bishop. *Pattern Recognition and Machine Learning*, Springer, 2007.
- [BL97] J.R. Birge and F. Louveau. *Introduction to Stochastic Programming*, Springer, 1997.
- [Bovik05] A. Bovik (ed.). *Handbook of Image & Video Processing*, Elsevier Academic Press, 2005.
- [BTN02] A. Ben-Tal and A. Nemirovski. "Robust optimization – methodology and applications." *Mathematical Programming*, 92(3):453-480, 2002.
- [CBM+06] H.N. Chapman, A. Barty, S. Marchesini, A. Noy, S.P. Hau-Riege, C. Cui, M.R. Howells, R. Rosen, H. He, J.C. Spence, U. Weierstall, T. Beetz, C. Jacobsen and D. Shapiro. "High-resolution ab initio Three-dimensional X-ray Diffraction Microscopy." *Journal of the Optical Society of America A: Optics, Image Science, and Vision*, 23(5):1179-1200, 2006.
- [CES+13] E.J. Candès, Y.C. Eldar, T. Strohmer and V. Voroninski. "Phase Retrieval via Matrix Completion." *SIAM Journal on Imaging Sciences*, 6(1):199-225, 2013.
- [CL07] T.M. Cioppa and T.W. Lucas. "Efficient nearly orthogonal and space-filling Latin hypercubes." *Technometrics*, 49(1):45-55, 2007.
- [DXF+12] F. De Carlo, X. Xiao, K. Fezzaa, S. Wang, N. Schwarz, C. Jacobsen, N. Chawla and F. Fusses. "Data Intensive Science at Synchrotron Based 3D X-ray Imaging Facilities." 8th IEEE International Conference on E-Science, Chicago, Illinois, October 8-12, 2012, eScience, pp. 1-3, 2012.
- [EK12] Y.C. Eldar and G. Kutyniok. *Compressed Sensing: Theory and Applications*, Cambridge University Press, 2012.

[Frank06] J. Frank. *Three-Dimensional Electron Microscopy of Macromolecular Assemblies: Visualization of Biological Molecules in Their Native State*, Oxford University Press, 2006.

[GCW+09] I. Gorton, J. Chase, A. Wynne, J. Almquist and A. Chappell. "Services + Components = Data Intensive Scientific Workflow Applications with MeDICI." *Component-Based Software Engineering, Lecture Notes in Computer Science*, 5582:227-241, 2009.

[Hansen97] P.C. Hansen. *Rank-Deficient and Discrete Ill-Posed Problems*, SIAM, 1997.

[HHH+07] S. Habib, K. Heitmann, D. Higdon, C. Nakhleh and B. Williams. "Cosmic calibration: Constraints from the matter power spectrum and the cosmic microwave background." *Physical Review D*, 76(8):083503, 2007.

[HR10] H.R. 5116—111th Congress: America COMPETES Reauthorization Act of 2010. (2010). In: www.GovTrack.us. Last retrieved July 22, 2013 from <http://www.govtrack.us/congress/bills/111/hr5116>.

[HTT09] T. Hey, S. Tansley and K. Tolle (eds.). *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft Research, 2009.

[IES+11] P. Isenberg, N. Elmqvist, J. Scholtz, D. Cernea, K-L Ma and H. Hagen. "Collaborative visualization: Definition, challenges, and research agenda." *Information Visualization*, 10(4):310-326, 2011.

[KBB+08] P. Kogge, K. Bergman, S. Borkar, D. Campbell, W. Carlson, W. Dally, M. Denneau, P. Franzon, W. Harrod, K. Hill, J. Hiller, S. Karp, S. Keckler, D. Klein, R. Lucas, M. Richards, A. Scarpelli, S. Scott, A. Snavey, T. Sterling, R.S. Williams and K. Yelick. "ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems." DARPA IPTO Whitepaper, 2008. Available online at: <http://www.cse.nd.edu/Reports/2008/TR-2008-13.pdf>.

[LHL08] M.E. Liggins, D.L. Hall and J. Llinas (eds.). *Handbook of Multisensor Data Fusion: Theory and Practice, Second Edition*, CRC Press, 2008.

[Lipeles12] E. Lipeles. "L1 track triggers for ATLAS in the HL-LHC." *Journal of Instrumentation*, 7:C01087, 2012.

[LZL+10] J. Lofstead, F. Zheng, Q. Liu, S. Klasky, R. Oldfield, T. Kordenbrock, K. Schwan and M. Wolf. "Managing Variability in the IO Performance of Petascale Storage Systems." In: *SC'10: Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*, New York, N.Y., 2010.

[MCK+99] J. Miao, P. Charalambous, J. Kirz and D. Sayre. "Extending the methodology of X-ray crystallography to allow imaging of micrometre-sized non-crystalline specimens." *Nature* 400:342-344, 1999.

[MHC+03] S. Marchesini, H. He, H.N. Chapman, S.P. Hau-Riege, A. Noy, M.R. Howells, U. Weierstall and J.C.H. Spence. "X-ray image reconstruction from a diffraction pattern alone." *Physical Review B*, 68(14):14010(R), 2003.

[Modersitzki04] J. Modersitzki. *Numerical Methods for Image Registration (Numerical Mathematics and Scientific Computation)*, Oxford University Press, 2004.

- [MSF+10] B. Matthews, S. Sufi, D. Flannery, L. Lerusse, T. Griffin, M. Gleaves and K. Kleese van Dam. "Using a Core Scientific Metadata Model in Large-Scale Facilities." *International Journal of Digital Curation*, 5(1):106-118, 2010.
- [NWV+00] R. Neutze, R. Wouts, D. van der Spoel, E. Weckert and J. Hajdu. "Potential for biomolecular imaging with femtosecond X-ray pulses." *Nature*, 406(6797):752-757, 2000.
- [RBU08] S. Ramani, T. Blu and M. Unser. "Monte-Carlo Sure: A Black-Box Optimization of Regularization Parameters for General Denoising Algorithms." *IEEE Transactions on Image Processing*, 17(9):1540-1554, 2008.
- [ROF92] L.I. Rudin, S. Osher and E. Fatemi. "Nonlinear total variation based noise removal algorithms." *Physica D*, 60(1-4):259-268, 1992.
- [RW05] C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning*, MIT Press, 2005.
- [SD04] J.C.H. Spence and R.B. Doak. "Single Molecule Diffraction." *Physical Review Letters*, 92(19):198102, 2004.
- [SF12] C.M. Schafer and P.E. Freeman. "Likelihood-Free Inference in Cosmology: Potential for the Estimation of Luminosity Functions." In: *Statistical Challenges in Modern Astronomy V, Lecture Notes in Statistics*, pp. 3-189, Springer, 2012.
- [SPK13] E.G. Stephan, P. Pinheiro and K. Kleese van Dam. "Bridging the Gap between Scientific Data Producers and Consumers: A Provenance Approach." In: *Data-Intensive Science*, pp. 279-300, CRC Press, 2013.
- [WDM+01] Y. Wang, F. De Carlo, D.C. Mancini, I. McNulty, B. Tieman, J. Bresnahan, I. Foster, J. Insley, G. von Laszewski, C. Kesselman, M-H Su and M. Thiebaux. "A high-throughput x-ray microtomography system at the Advanced Photon Source." *Review of Scientific Instruments*, 72(4):2062-2068, 2001.
- [WFI+09] M. Wilde, I. Foster, K. Iskra, P. Beckman, Z. Zhang, A. Espinosa, M. Hategan, B. Clifford and I. Raicu. "Parallel Scripting for Applications at the Petascale and Beyond." *Computer* 42(11):50-60, 2009.
- [WYL12] Z. Wen, C. Yang, X. Liu and S. Marchesini. "Alternating direction methods for classical and ptychographic phase retrieval." *Inverse Problems*, 28(11):115010, 2012.
- [Youngman12] C. Youngman. "Data Acquisition and Controls." *European XFEL Users' Meeting*, January 25-27, 2012, Hamburg, Germany. Available online at: http://www.xfel.eu/sites/site_xfel-gmbh/content/e63594/e65073/e126274/e134393/3Youngman_DataAcquisitionandControls_eng.pdf.



9 Participants and Contributors

Computer Science and Applied Math Scientists

Last Name	First Name	Lab/University
Ahrens	Jim	LANL
Bethel	E. Wes	LBNL
Choudhary	Alok	NWU
Foster	Ian	ANL
Geist	Al	ORNL
Hendrickson	Bruce	SNL
Higdon	Dave	LANL
Kamath	Chandrika	LLNL
Klasky	Scott	ORNL
Kleese van Dam	Kerstin	PNNL
Leyffer	Sven	ANL
Li	Xiaoye (Sherry)	LBNL
Ma	Kwan-Liu	UC-Davis
Pascucci	Valerio	Utah
Rogers	David	SNL
Ross	Rob	ANL
Shipman	Galen	ORNL
Shoshani	Arie	LBNL
Skinner	David	LBNL
Wilson	Alyson	IDA
Yang	Chao	LBNL
Yu	Dantong	BNL

Domain Scientists

Last Name	First Name	Lab/University
Boehnlein	Amber	SLAC
Habib	Salman	ANL
Jacobsen	Doug	LBNL
Parkinson	Dula	LBNL
Prendergast	David	LBNL
Proffen	Thomas	ORNL
Roser	Robert	Fermilab
Tull	Craig	LBNL
Williams	Dean	LLNL



Meeting Organizers

Bruce Hendrikson, Arie Shoshani

Breakout Sessions Leaders

Data Processing: Kerstin Kleese van Dam, Ian Foster

Data Management: Rob Ross, Al Geist, Galen Shipman

Data Analysis: Chandrika Kamath, Jim Ahrens

Use Case Providers

Basic Energy Sciences

- Light Sources: Dula Parkinson
- Nanoscience Centers: David Prendergast
- Neutron Facilities: Thomas Proffen

Biological and Environmental Research

- Climate Change Research: Dean Williams
- Genome Science: Douglas Jacobsen and Dantong Yu
- Environmental Molecular Sciences: Kerstin Kleese van Dam

High Energy Physics

- Energy Frontier: Robert Roser
- Intensity Frontier: Craig Tull
- Cosmic Frontier: Salman Habib



10 Acknowledgments ---

The workshop organizers wish to thank Lucy Nowell and Ceren Susut for facilitating the meeting and soliciting contributions on scientific drivers from domain scientists working with the BER, BES, and HEP offices. The organizers also wish to acknowledge Carolyn Lauzon of DOE for her help in collecting and preparing the acronyms descriptions.