



DOE Data Day 2019 Report

September 25–26, 2019

Livermore, CA, USA

Convened by

Lawrence Livermore National Laboratory (LLNL)
on behalf of U.S. Department of Energy (DOE) laboratories

Organizing Committee

Jessie Gaylord (LLNL)

Ghaleb Abdulla (LLNL)

Daniel Laney (LLNL)

Stanley Ruppert (LLNL)

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

Lawrence Livermore National Laboratory is operated by Lawrence Livermore National Security, LLC, for the U.S. Department of Energy, National Nuclear Security Administration under Contract DE-AC52-07NA27344. LLNL-TR-799308.

Contents

EXECUTIVE SUMMARY	3
ACKNOWLEDGMENTS	4
INTRODUCTION AND EVENT MOTIVATION	5
WHITE PAPER	5
AGENDA AND ABSTRACTS	8
SESSION 1: DATA CURATION AND STANDARDS	8
SESSION 2: DATA-INTENSIVE COMPUTING	10
SESSION 3: DATA MANAGEMENT IN THE CLOUD	12
SESSION 4: DATA ACCESS, SHARING, AND SENSITIVITY	13
POSTERS	15
CONCLUSION AND RECOMMENDATIONS	19
APPENDICES.....	21
ORGANIZING COMMITTEE	21
ATTENDEES.....	23
SURVEY RESULTS	31
ACRONYMS.....	34

Executive Summary

The Department of Energy (DOE) has joined the larger scientific community in the promotion of data management as a means to higher quality, more efficient research and analysis, and as a critical component of data science. Tools and platforms to support data management and analysis are rapidly evolving and provide enormous opportunities. They also pose challenges that can be specific to DOE but are common across DOE mission areas and organizations.

The DOE Data Day (D3) workshop was held on September 25–26, 2019, and organized by a team at Lawrence Livermore National Laboratory (LLNL) in an effort to gather the data management practitioners at the DOE labs together to share their work and results, facilitating knowledge transfers and best practices across project teams. Over 100 attendees participated from multiple national labs—LLNL, Sandia (SNL), Lawrence Berkeley (LBNL), Los Alamos (LANL), Oak Ridge (ORNL), Argonne (ANL), Pacific Northwest (PNNL), Stanford Linear Accelerator (SLAC)—along with Mission Support and Test Services (MSTS), National Energy Technology Laboratory (NETL), Pantex, National Security Complex (Y-12), National Nuclear Security Administration (NNSA), the DOE Office of Scientific and Technical Information (OSTI), and elsewhere in the DOE complex.

A call for abstracts was distributed via email to people who had previously expressed an interest in the concept during informal and ad hoc meetings with organizers as well as to people with known involvement in data management at the national labs. The response was overwhelmingly supportive, and almost 50 abstract submissions were received. Themes emerged from the abstract submissions, so sessions were organized into four topics:

- Data Curation and Standards
- Data-Intensive Computing
- Data Management in the Cloud
- Data Access, Sharing, and Sensitivity

These subject areas provided the framework for the agenda, which featured talks given by SNL, LLNL, NETL, SLAC, OSTI, PNNL, ANL, ORNL, and LBNL researchers. These topics were also the basis for a half-day of breakout discussions followed by verbal reports from the breakout teams. To give more participants the opportunity to share their perspectives, a poster session was held and the 21 poster presenters each gave two-minute lightning talks. The agenda also included a group photo, a no-host dinner, and optional tours of LLNL's National Ignition Facility (NIF) and high-performance computing (HPC) facility.

Participant discussions and engagement during the workshop were phenomenal. There was a clear consensus that the workshop should become a series, and other DOE labs volunteered to host and help organize future D3 events. An NA-122 PRIDE (Product Realization Integrated Digital Enterprise) meeting organized by DOE's Kansas City National Security Campus was moved from Kansas City to LLNL and held the day before D3 to take advantage of expected synergies between the events. PRIDE organizers also expressed interest in future collaborations with D3.

This Report will summarize the important discussions and recommendations from the different working sessions and contains the agenda, submitted abstracts, posters, and list of registered attendees. The Report

will be distributed to DOE, each participating institution's programmatic stakeholders, and attendees. A [dedicated D3 website](#) will link to the Report, presentation slides, poster files, and other materials associated with the event. The website will also host future planning information.

Acknowledgments

The D3 workshop was made possible by funding from the Nonproliferation Research and Development (NA-22) data science portfolio, significant administrative support from LLNL's Weapons and Complex Integration directorate (WCI), and the efforts of the D3 organizing committee led by Jessie Gaylord and including Dan Laney, Stan Ruppert, Ghaleb Abdulla, and Loni Hoellwarth. The organizing team represents global security, basic science, and weapons complex missions at LLNL.

In particular, NA-22 Data Science Program Manager Angie Waterworth provided funding for planning, abstract reviews, refreshments, photography, and this Report. WCI provided funding for administrative support, hospitality, and the [event registration website](#). We appreciate Elizabeth Brown's, Angeline Lee's, and the PRIDE organization's flexibility in the spirit of the event's multi-sponsor, multi-mission, and multi-domain relevance.

The organizers would like to extend special thanks to the LLNL administrative team who coordinated the event including meals, venue and hotel arrangements, onsite transportation, visitor badging, and facility tours. Loni Hoellwarth, Adilene Cuevas, Terri Stearman, and Jacqueline Jolley managed these event logistics and ensured participant safety.

Introduction and Event Motivation

Data is critical to all DOE work. Data management encompasses many activities and considerations—curation, extraction, storage, preservation, tracking, access, security, transfer, retrieval, and more—for a wide range of data formats and quality. It requires a disciplined approach to metadata, which tracks data provenance and provides traceability from raw data products through analysis results and potentially through production.

The inaugural DOE Data Day workshop, abbreviated to D3, was born from this critical work and held on September 25–26, 2019. The original day-long agenda was expanded to two days due to enthusiastic response to the call for poster and presentation abstracts. Hosted by LLNL, the event welcomed more than 100 participants from across the DOE complex.

D3's primary goals were to bring DOE institutions together to share their data management use cases, challenges, and solutions; identify potential synergies and efficiencies; and establish proactive channels for future collaborations. The event crossed program boundaries and mission areas, with participants exploring best practices and the latest technologies to help DOE researchers leverage new techniques, respond to data security threats, and advance fundamental science in valuable ways.

After 18 presentations and 21 posters, plus LLNL facility tours and working group breakout sessions, the event was deemed a success. Participant feedback indicated a strong preference for making D3 an annual event, as it fills a void not met by existing venues (e.g., domain specific, commercial or revenue driven, academic/open data). Ultimately, D3 helps raise the bar on how valuable DOE data assets are and can be managed.

White Paper

LLNL co-organizer Jessie Gaylord submitted this white paper to propose the idea to sponsors, outline the event's purpose to organizers and stakeholders, and generate interest at other laboratories.

Proposal Summary

Most programs at the national labs either generate data, are wholly dependent on the availability of data, or both. For these programs data management supports transparency, collaboration, and a higher overall return on research and development investments. To support this increasing laboratory resources are invested in developing data ingestion and curation systems across all mission spaces, but often these efforts exist in programmatic stovepipes. This proposal is for a one-day conference for national laboratory data managers and system developers to share technologies and solutions with the goal of lowering the learning curve for new projects, improving consistency in how data is handled across the complex, and developing best practices.

Need

DOE has joined the larger scientific community in the promotion of data management as a means to higher quality and more efficient research. Data management includes a disciplined approach to metadata which tracks provenance and provides traceability from raw data products to analytic results. Effective curation ensures long term data access and security. Together metadata and curation support repeatability,

attribution, improved research quality, collaboration, and transparency. In addition, the rise of data-driven modeling, artificial intelligence, and machine learning (ML) is forcing changes in laboratory data centers in order to integrate experimental data with large computational data sets. Novel approaches and systems are required to meet data management goals and ensure data assets are available to future researchers working on broader science questions than today's.

Current State of the Art

Numerous organizations have formed to service the growing need for data management in a world increasingly driven by data. An ever-wider variety of commercial and open-source software is available for data processing and curation, and the global call for reproducible research in science communities is fostering new tools for packaging data and software into reproducible artifacts. Organizations such as NSF and NASA sponsor multiple projects with online platforms, publications, and educational venues for increasing data management awareness and developing data standards in research communities.

While scientific and commercial entities provide important educational resources and solutions for data management practitioners, they are blind to key aspects of national laboratory work that have significant implications on data management. Scientific data organizations are usually specific to particular research domains and do not cover all aspects of national security. They are also frequently targeted to academia and dedicated to the principles of open science which do not translate well to the closed networks and sensitive data at the national labs. Commercial and open-source data solutions are primarily geared towards business applications and may not support laboratory workflows or cyber security requirements without considerable customization.

Many lab-specific data management challenges are due to high dependencies on legacy and sensitive data, data that is very expensive to generate or cannot be reproduced, historically owner-based data management practices and cultures, and specialized cyber security policies. Consequently there is not a clear venue for national labs to discuss the particular challenges of developing standards-based processes and systems to manage volumes of national security data in lab environments. Since data management is a support function for other work, cross-program and cross-lab conversations happen as an add-on in the context of other topics, in infrequent and narrowly scoped technical exchanges between individual practitioners, or not at all.

Proposed Approach

A one-day data conference dedicated to data management work at the DOE national labs (possibly named DOE Data Day, or "D3" for short) would provide an extremely valuable forum for data management practitioners and system developers. Many programs are investing more formally in data management, and open discussions promoting shared solutions and best practices that are effective in lab environments are critical to make efficient progress in this fast-moving field. Presentations and discussions on data, software, storage, and network topics specific to lab programs and constraints would be enormously valuable to multiple missions. Potential topics include:

- Metadata standards for diverse datasets
- Particular challenges of legacy data and missing metadata
- Data pipeline software and methods
- Data infrastructures for analytics

- Data sharing across isolated networks and between labs
- Moving, managing, and storing large volumes of data
- Commercial cloud usage at the labs
- Managing sensitive data
- Multi-Lab authentication, cyber approvals, and other data security considerations
- Data archiving, processing, and sharing on classified networks
- Leveraging experimental and large scale simulation data for analysis and discovery

Developers, data managers, data generators, and IT support personnel at the national labs would be encouraged to participate in this event. Presentations would highlight developing approaches and effective existing solutions in a variety of scientific domains. Informal or organized discussions would facilitate information sharing, collaborations, and better integrations between programs. The objective of the conference is to promote awareness of effective data management strategies, shorten the learning curve for new efforts, and increase the overall quality of data management practice at the national labs.

Repeat events may be planned for future years based on interest. Over 20 personnel from multiple national labs (LLNL, Argonne, PNNL, SNL, ORNL, and LANL) and representing diverse programs and mission spaces have already asked to participate in the first D3 through informal discussions and word-of-mouth. Funding is being solicited from interested sponsors in climate security, nonproliferation, and defense with the anticipation that costs might be shared. This will be a highly collaborative effort with broad potential impact across the DOE laboratory complex.

LLNL-PROP-763984 (applies to White Paper only)

Agenda and Abstracts

The D3 workshop was organized into four sessions with three to six speakers in each: Data Curation and Standards; Data-Intensive Computing; Data Management in the Cloud; and Data Access, Sharing, and Sensitivity. Speakers were allotted 20 minutes for their presentations plus a five-minute question-and-answer period. In addition to these workshop sessions, the agenda including working group breakout sessions; poster lightning talks; and tours of NIF and LLNL's largest machine room, home of the Sierra supercomputer. Abstracts are summarized here. See the [event registration website](#) for full abstracts, presentation slides, and poster files.

Session 1: Data Curation and Standards

The first D3 session featured speakers from six DOE organizations who focused on data curation and standards for a range of related activities—access control, data formats, documentation, and more. Workshop participants heard about custom data management systems, lessons learned during platform upgrades and implementations, and ongoing data preservation efforts. Also discussed were the DOE's Data Identification Service and best practices for establishing and enforcing data management standards.

Citadel: Stockpile Evaluation Data System

William DeRaad | SNL

William DeRaad from SNL presented the Stockpile Evaluation Data System (SEDS)— a next-generation data repository that enables centralized access to stockpile evaluation data for the life of the system. SEDS has been architected to enable modern analytics practices while still providing value from a history of legacy test data. The primary SEDS use case is to support stockpile evaluation data users who use the data to evaluate stockpile safety and performance. Unlike development test data repositories, this product archives official record copy to support stockpile assessment and decisions, and provides the provenance for the test data for the life of the system, which supports a strategy to have long-life, multi-use data available and accessible to all who need it and have the need to know (NTK). DeRaad shared details about the SEDS system's architecture and his team's methodologies that address data provenance, metadata standards, and requirements analysis.

The Earth System Grid Federation: Management of Distributed Data

Sasha Ames | LLNL

Sasha Ames of LLNL described the Earth System Grid Federation (ESGF), which is a collaboration that develops, deploys, and maintains software infrastructure for the management, dissemination, and analysis of climate model output and observational data. The ESGF has addressed the problem of making large-scale data accessible in a distributed fashion when such data originates at geographically disparate sources worldwide (e.g., computing centers that run climate models or sites that curate collections of observation). For example, each participating institution hosts the data, but indexing services are handled remotely by select index sites. The ESGF faces challenges in the deployment of a system that incorporates distributed data, federated identity services, and replicated, distributed indexing. A particular challenge addressed at LLNL has been the automated process of replicated data, needed to enhance data availability for the worldwide community; LLNL serves as the leading center with the largest archive, which is expected to grow

to 11 petabytes and beyond. Looking ahead, Ames explained, the ESGF team plans to refresh the system's architecture and constituent components and is exploring Cloud-replicated, distributed services.

Lessons Learned from Present and Past Data Preservation Efforts to Build "Smart" Tools for Fossil Energy Data Products

Kelly Rose | NETL

Kelly Rose from NETL introduced the Energy Data eXchange (EDX), a virtual data library and laboratory that manages data for gas shales, carbon capture and storage, materials, and other energy and geoscience projects. Following a 2013 *Nature* study that documented loss of over 80% of scientific data underlying journal publications, a growing and persistent shift exists in the value and importance place on data products derived from research and development. The value of research data products has increased, gaining recognition as significant products worthy of digital object identifiers (DOIs) and citations of their own to accompany more conventional publication- and presentation-related research products. Like the broader community, Rose noted, DOE fossil energy researchers and stakeholders are seeking access to both present-day data products as well as historical data products. The EDX team regularly field requests for assistance in finding and connecting to federally funded data products from the last several decades, predating the modern era of data repositories. NETL has had success in tracking down a variety of those resources and ensuring their preservation.

Overview of the LCLS Data Management System

Amadeo Perazzo | SLAC

The Linac Coherent Light Source (LCLS) is an x-ray free electron laser user facility at SLAC. In its 10-year history, LCLS has collected several petabytes of data serving thousands of users across a wide spectrum of science. LCLS provides the resources required to collect, process, and store the data generated by the experimenters. In this presentation, Amadeo Perazzo offered an overview of the system including data collection, annotation of the data with metadata, and the flow of the data through the different storage resources (e.g., transfers to external sites). The presentation also described how users interact with LCLS and process and manage their data. The data movement and underlying storage technologies will be presented in some detail. A new accelerator, LCLS-II, will increase x-ray pulse rates and collect even more data faster. Perazzo explained how his team is planning to accommodate these new requirements.

Assigning DOIs to Research Data Through the DOE Data Identification Service

Sara Studwell | OSTI

OSTI provides a free service for DOE-funded research by assigning DOIs to datasets and data collections. Assigning DOIs to research data through the DOE Data Identification (ID) Service helps facilitate citation, discovery, retrieval, and reuse of data. Sara Studwell explained how OSTI works with data managers and producers to define metadata and build intelligence into the DOI itself. The DOI resolves to a landing page that describes and links to the data hosted by the laboratory or facility. Data records, including the DOIs and associated metadata, are included in the osti.gov and DOE Data Explorer search tools, which are indexed by Google, Google Dataset Search, and other search engines, raising the data's discoverability and visibility. Studwell, an OSTI librarian, also described the online DOI assignment process wherein researchers complete required and optional fields (some with controlled vocabularies) and can cross-link their research results to related DOIs.

Making Data Standards Work for You: Leveraging Community Standards and Best Practices to Support Your Scientific Research and Data Management Practices

Eric Stephan | PNNL

The application of community-inspired, FAIR—findable, accessible, interoperable, reusable—standards increases not only the lifespan of data products, it also preserves the intrinsic value of present and future research investments. FAIR is built upon many stable standards communities such as the International Engineering Task Force, the International Standards Organization, Object Management Group, World Wide Web Consortium, the Hierarchical Data Format Group, the National Institute of Standards and Technology, and the Research Data Alliance. Eric Stephan of PNNL discussed how to incentivize and encourage the research community to make standards-based approaches a key part of planning and implementation of scientific research, thus reducing research and development costs and providing a means to extend data lifespan of beyond the research endeavor. Stephan also provided an overview of the standards landscape and how it has been applied to the DOE/Advanced Scientific Computing Research (ASCR) program’s Resource Discovery for Extreme Scale Collaborations for bridging curated soil moisture metadata, reproducibility to support Energy Exascale Earth System Model simulations, and Smart Grid interoperability.

Session 2: Data-Intensive Computing

The second D3 session showcased data management ecosystems at multiple labs, explored the challenges data collection and analysis at world-class, high-throughput experimental facilities. Additional presentations explored data complexities—visualization, analysis, interdependencies, abstractions—in the context of exascale computing systems.

Role of Data Curation at the National Ignition Facility

Philip Adams | LLNL

Data is critical to NIF’s success as the world’s largest laser facility, explained LLNL’s Philip Adams. From the systems that track the machine configuration, the systems that control the laser, and the systems that analyze the experimental results, the data is varied as much as the sources are diverse. Images, sensor data, database rows and columns all need to be aggregated and applied towards solving problems, discerning patterns, and identifying opportunities. While much coverage about Big Data focuses on artificial intelligence, the Internet of Things, and creating data lakes, Adams focused on the role of data curation in Big Data. Good data management practices are essential for ensuring that research and operational data are of high quality, accessible, and sustainable for the long term. The goal of data curation is to ensure that data can be retrieved for future research and/or trend analysis in the most cost-effective way. Adams also reviewed the decisions made in handling NIF data and summarized future directions.

Globus Research Data Platform

Rick Wagner | ANL

Research data management challenges include varied data formats, analysis on distributed resources, catalogs of metadata, and dynamic collaborations around analysis. Rick Wagner from ANL presented an overview of Globus (globus.org), a research platform developed by ANL and operated by the University of Chicago. Globus’s services are widely used within and outside the DOE, with tens of thousands of users and more than 14,000 storage systems at leading U.S. universities and research computing centers. Globus provides high-performance, secure file access, transfer, and synchronization directly between storage

systems (i.e., without needing to relay via an intermediary machine). The platform scales to meet the needs of increasingly diverse data by handling all the difficult aspects of data transfer—from authentication at source and destination to performance optimization and automatic fault recovery. This platform-as-a-service integrates with other systems to handle transfer and sharing capabilities into scientific Web applications, portals, and elsewhere. Wagner described Globus’s core components and advanced features that, for instance, ensure that users who access shared endpoints are restricted to the locations and permissions granted by the owner.

Firebird: Integrated Multi-Phenomenology Data and Analytics Platform

Elaine Martinez | SNL

Within the Department of Defense and intelligence community, discovery and exploitation of multi-phenomenology information is currently a difficult challenge. As the complexity of the problem grows, the inadequacies of the current computing and data systems becomes more evident: They were simply not designed for analysis across multiple phenomenologies and often struggle with storing and accessing large volumes of data. SNL has combined decades of knowledge and experience in software and hardware architectures, cybersecurity, sensor data, and Big Data analytics to create a Cloud-based architecture that supports these complex problems. SNL’s Elaine Martinez outlined her team’s methodology for implementing Firebird 1.0, a data agnostic and reusable architecture that ingests multi-phenomenology data—in various data formats and from a variety of sources. Martinez also described Firebird 2.0, which will leverage new industry standards in open-source technologies and processes as well as modernize the way analysts and scientists can exploit disparate, multi-phenomenology data to better serve the intelligence community.

Infrastructure for Managing Scientific Data at the Exascale

Kshitij Mehta and Lipeng Wan | ORNL

The advent of exascale computing has led to the emergence of novel supercomputer architectures and new classes of simulations. Modern heterogeneous supercomputer architectures utilize a deep, complex memory and storage hierarchy. Consequently, simulations increasingly focus on improved methods of managing large data. Multiphysics codes, in situ workflows, and ensemble runs have introduced new challenges in data management, as data generated by these applications exhibit properties of all five V’s of Big Data: volume, velocity, variety, veracity, and value. ORNL is home to Summit, the world’s fastest supercomputer. ORNL’s Kshitij Mehta and Lipeng Wan discussed the challenges faced by science applications that utilize the high-performance ADIOS ecosystem, which is an ORNL-developed middleware library that enables data management, analysis, and visualization. Mehta and Wan outlined their team’s approach to solving two issues: the need for a modern metadata format for self-describing data, and the need for in situ workflow tools for dynamic management of applications and data. The team built a metadata mechanism called BP4 that inherits ADIOS’s self-describing file format and significantly reduces the metadata overhead.

Fusion of Big Data and Traditional Visualization Tools and Workflows

Mark Miller | LLNL

Visualization tools such as LLNL’s VisIt, ParaView, and EnSight have been successful in HPC despite many challenges their respective development teams face—such as relatively high costs to develop and maintain, complexities in implementation due to a myriad of scientific data models and storage paradigms they must support, ever-widening varieties of disparate software technology underpinnings, inflexibilities in data

parallelization and query scope, and challenges in hiring and retaining software engineering staff with the necessary expertise. Big Data technology, appropriately leveraged within these tools, can potentially address many of these issues. Using VisIt as an example, LLNL's Mark Miller explained how these issues manifest and what Big Data technology can do to address them. His team developed a key-value approach for representing scientific data models in a Big Data-friendly way, enabling natural integration with existing visualization tool front-ends. Miller also proposed a new hybrid HPC/Big Data architecture for VisIt and demonstrated preliminary concepts in performance and capability enablement as well as software engineering cost reduction.

Proactive Data Containers: An Intelligent Object-Centric Data Management System for HPC Suren Byna | LBNL

Parallel file systems face fundamental challenges in scalable metadata operations, semantics-based data movement performance tuning, and asynchronous operation. Furthermore, storage systems on upcoming exascale supercomputers are being deployed with an unprecedented level of complexity due to a deep system memory/storage hierarchy-based architectures. Suren Byna of LBNL presented a user-level, object-centric data management system called Proactive Data Containers (PDC), which provides abstractions and storage mechanisms that take advantage of deep memory and storage hierarchy, enable proactive automated performance tuning in storing and retrieving data, and perform user-defined analytics in the data path on large-scale supercomputing systems. In the PDC system, scalable metadata management is achieved using the memory available in compute nodes. discuss automatic data analysis and transformations while the data is moving from one location to another. Byna also described the PDC concepts of automatic reorganization and placement of data in the memory and storage hierarchy, closer to data analysis using the history of previous data accesses for analysis and of any user-provided hints.

Session 3: Data Management in the Cloud

In D3's third session, speakers discussed the opportunities available with Cloud computing architectures. For example, the ability to spin up and decommission resources on demand is attractive to programs seeking scalable storage, elastic compute, and fast start-ups. These presentations dived into several major considerations such as choosing appropriate platforms, ensuring security, and controlling costs. As Cloud technology advances, addressing the role of Cloud-based data management in labs' infrastructure is an important and developing effort.

Ecosystem for Open Science Tammie Borders | INL

As Tammie Borders from Idaho National Laboratory (INL) explained, the Ecosystem for Open Science (eOS) leverages commercial technologies to build a Cloud-based collaborative platform aimed at improving Defense Nuclear Nonproliferation (DNN) research and development activities through openness and sharing of data and information across projects. The outcome will improve data analysis, archiving, and disposition. Current data management challenges include specialized data formats, stove-piped data management practices with access limited to principal investigator or home institution, lack of standardized data management practices across projects to provide storage and access for long-term analysis, lack of standardized analytics capability, and potential for lapse in data service due to funding gaps and complex funding models. Implementation of the eOS improvements will spur a cultural shift in DNN research and development, potentially leading to

accelerated discoveries and new cross-project research while advancing the development of technical capabilities at national labs. Borders also described eOS requirements including Cloud deployment, single sign-on with multi-institutional access (i.e., a single access point), a searchable file repository, collaboration tools, and an analytics workspace.

Mission Enterprise in the Cloud

Katie Knobbs and Clay Hagler | PNNL

Katie Knobbs and Clay Hagler outlined PNNL's new multi-organization, Cloud-based collaboration and application hosting environment. With over 1,000 users across 45 federal, state, and local organizations, the mission-focused enterprise serves NNSA's Nuclear Incident Response mission, providing the DOE and its partners a common workspace for real-time communications, data, and files—thus enabling users from various organizations to collaboratively plan, conduct, and assess radiological emergency response operations. The new platform includes account management, invoicing/billing, cybersecurity, application hosting, operations/monitoring, and user support. Knobbs and Hagler reviewed the project's challenges, such as adhering to multiple sets of cybersecurity requirements, providing mission-ready authentication and common credential solutions, and streamlining account management and approvals. The team also had to onboard 650 users in fewer than 9 months and provide training in new collaboration tools. The experience gave the team valuable insights into the needs of the Nuclear Incident Response Office and the game-changing technology that can be developed and configured to support those needs.

An Extensible, Reusable Hybrid Cloud Data Management Platform

Chitra Sivaraman | PNNL

Another PNNL speaker, Chitra Sivaraman, described a state-of-the-art data management platform based on a hybrid Cloud architecture and industry best practices. The system provides a range of scientific data management features such as data collection, transport, storage, archival, security, curation, quality assurance, provenance tracking, metadata standards, processing pipelines, publication (DOIs), and metrics tracking. The team leverages Amazon Web Services (AWS) to simplify development and deployment, allow the application to scale on demand, and increase system reliability and uptime. The hybrid Cloud architecture reduces data costs for archival storage while still allowing data to be easily available in the Cloud for analytic pipelines. As Sivaraman explained, PNNL's framework also uses AWS lambda service hooks and dependency injection at the user interface as well as representational state transfer (REST) services layers to enable customization of metadata, security policy, storage, and other features. Sivaraman also highlighted three communities that are currently using the framework to manage public-unlimited rights to restricted-proprietary data.

Session 4: Data Access, Sharing, and Sensitivity

The DOE's security requirements for shared information are necessarily stringent, so the final D3 session concentrated on the challenges and possible solutions for appropriately handling sensitive and NTK data. Two labs shared their experiences with collaborative, access-controlled platforms built in-house, while another presentation introduced best practices for managing high-quality common reference data.

Information Technology Lessons Learned from Eight Years of NTK

Susan Byrnes | SNL

Built in the early 2000s, SNL's flexible and configurable data access control system for enforcing consistent NTK access control is still in use by several SNL applications today. Susan Byrnes described a project under way to adapt the existing system to meet today's evolved challenges. Project goals include ensuring scalability (i.e., to enable multiple instances to be deployed for performance and availability) and balancing flexibility and complexity with simplicity and transparency. Byrnes explained that SNL user expectations for the system have changed over time, mostly regarding physical versus electronic documents, processing larger volumes of data, and the need for instantaneous access. She also reviewed the system's existing features that are still beneficial as well as those that require updating, and outlined challenges related to sufficient security metadata, the lack of security information for legacy data, solutions for aggregated data, integration with role-based access control, and more. The lessons the team has learned will contribute to the success of the current system upgrades and future information technology (IT) implementations.

Piloting a Collaborative Data Management Platform at LANL

Martin Klein and Brian Cain | LANL

Martin Klein and Brian Cain from LANL shared their lessons learned while developing and testing a pilot platform for collaborative data management. Motivated by Office of Science and Technology Policy, LANL library staff kicked off the Nucleus Project in 2018 to investigate non-Cloud (due to LANL policy) solutions that could support internal and external collaboration; data integration, preservation, storage, sharing; and security compliance. The resulting platform, Nucleus, is based on a local installation of the Open Science Framework and is connected to productivity portals (e.g., ownCloud, a locally hosted sync and share storage solution; GitLab, an internal source code repository). The platform supports researchers' collaborative goals at various stages of the research lifecycle. By incorporating functions to submit research output to LANL's institutional review-and-release system, Nucleus helps streamline research workflows. Klein and Cain described other Nucleus features, such as providing an overview of assets involved in research collaboration. Anticipated benefits for LANL include research lifecycle tracking, a single point of data preservation, and a seamless method for compliance with LANL's review-and-release and security policies.

Master Data within the Nuclear Security Enterprise (NSE)

Gregory Orndorff | SNL

SNL's Gregory Orndorff concluded the presentations with an explanation of master data management (MDM), which consists of high-quality common reference data (e.g., person or organization data) that can be used by multiple applications. The need for an MDM initiative arose because of inconsistent, lower quality data preventing accessibility and analytics of information, and efforts began in 2011 under the federal PRIDE program. Beginning with fully defined data governance policies and procedures, data stewards were identified, and web services became available for consumption by applications across the NSE. Orndorff showed how the MDM structure evolved and described its anticipated path forward, highlighting challenges and successes along the way. SNL's master data team works under the auspices of the PRIDE Data Governance Board to ensure program quality. As Orndorff discussed, many benefits can be realized by the enterprise from applications choosing to consume master data—among them are happier users, easier integration/sharing of data with other applications, reduced costs, sound decisions, decreased risk, and increased compliance.

Posters

The D3 workshop asked poster authors to give two-minute lightning talks summarizing their work and drawing the audience's attention to their specific posters, which were displayed around the event room. The lightning talks were scheduled in two groups between the main sessions, and the agenda provided time for attendees to circulate around the room and talk with poster authors as desired. Poster files and associated lightning talk slides are available on the [event registration website](#).

1. Data Management between DOE ACTICI Lab Partners

David Henderson | Y-12

David Henderson from Y-12 introduced the ACTICI program (Advanced Computer Tools to Identify Classified Information), a partnership between DOE labs that assists and improves confidence in classification decisions. ACTICI challenges include the difficulty of sharing and analyzing classified data (e.g., training/testing data and metadata) securely between geographically disparate lab partners, which must be coordinated between subject matter experts and computational scientists while ensuring security of classified material. Ultimately, the program will provide a data infrastructure framework for DOE classification reviewers.

2. Lessons Learned, Best Practices, and Emerging Technologies of Energy Data Management

Chad Rowan | NETL

As NETL's Chad Rowan explained, the EDX data library curates fossil energy-funded research and development data while providing users with an online collection of data, capabilities, and resources that advance ongoing research. EDX supports the entire data lifecycle by coordinating historical and current data from a variety of sources to facilitate access to research that crosscuts multiple NETL projects/programs, providing external access to technical products and data published by NETL-affiliated research teams, and collaborating with a variety of organizations and institutions in a secure environment through EDX's Collaborative Workspaces.

3. Creating Data Standards for Modern Experimental and Observational Sciences

Oliver Ruebel | LBNL

LBNL's Oliver Ruebel described the Neurodata Without Borders: Neurophysiology (NWB:N) data standards project, which includes development of HDMF, a hierarchical data modeling framework for modern science data standards. HDMF separates the data standardization process is separated into three main components: (1) data modeling and specification, (2) data storage and input/output, and (3) data use and user application programming interfaces (APIs). Thanks to HDMF, NWB:N supports a broad range of data types: extra- and intracellular electrophysiology recordings, optical physiology, behavioral data, and results of common data processing pipelines.

4. Citadel: Searching Multi-Layered Many-Typed Record Metadata

Rebecca Levinson | SNL

Core data in SNL's SEDS repository exists in test records, which must be readily searchable. As Rebecca Levinson from SNL explained, the challenge lies in fulfilling complex requirements, such as searching metadata fields in the test record itself and making SEDS widely extensible within Citadel projects. The team's search solution mines the backend for data to display, allowing a range of users to conduct diverse searches and filter results from a custom interface.

5. *Citadel: Tracking Data Provenance and Synchronizing Across Multiple Instances*

Aaron Comen | SNL

As presented by Aaron Comen, SNL's Citadel data framework leverages a graph database to track data provenance—that is, data sources and revision history for all data collected, even for data from isolated instances. Citadel stores all data and associated metadata as records as well as the relationships between revisions, using the latter to construct a directed acyclic graph representing a record's revision history. This solution simplifies synchronizing data across isolated instances and allows users to easily access old revisions.

6. *Data Frames: An Architecture for Managing Legacy Data*

Stephen Jackson | SNL

Managing data from legacy systems poses a complex challenge for a modern data system. As Stephen Jackson described, SNL has designed a portion of the Citadel framework to provide a convenient, partitionable, reproducible, binary format that consumes and modernizes legacy data while maintaining the integrity and provenance of the ingested data. This Data Frames architecture helps Citadel identify and extract data from legacy formats and provides API access to end-user analysis tools to allow them to load and manipulate data.

7. *Citadel: Dynamics Records*

Malachi Tolman | SNL

Due to countless permutations of user-defined data structures, SNL's Citadel team sought a flexible solution that would allow a customer to determine their own data structure before storing it in the institutional data system. Record schemas (objects that determine the structure of associated records) and dynamic records are used for organizing metadata and provide a consistent set of searchable metadata associated with each set of test data that logically goes together. Malachi Tolman noted that SNL is working to enforce standards on data structure and best practices.

8. *Frameworks for Explainable Artificial Intelligence*

Tammie Borders | INL

A scientific ecosystem for the integration of physics-based and ML models must be built to reach the goal of explainable artificial intelligence. This requires generalized data structures, robust analysis in data preprocessing, inference-based ML techniques, and visualization for ease of interpretation. As INL's Tammie Borders illustrated, applications of this framework include estimation of intrinsic kinetic information for reaction-diffusion problems and prediction of failure mechanisms for lithium-ion batteries. These approaches enable research to be targeted at specific areas of interest while improving understanding of poor performance of the objective.

9. *Scalable Coupled Analysis Workflow Support*

Kshitij Mehta and Lipeng Wan | ORNL

ORNL's Kshitij Mehta and Lipeng Wan described the challenges of managing data from large-scale simulations and making it flow efficiently to analysis pipelines. They highlighted one of ORNL's Exascale Computing Project applications—fusion whole device modeling in which scientists construct larger scientific computational experiments from the coupling of multiple individual fusion applications and run them in a coupled way. The team has developed a framework that leverages in-memory communications, HPC

workflow scheduling, and continuous running of coupled analysis, ultimately accelerating the time between a run and analysis of the science content.

10. Community Use of Persistent Sample Identifiers and Metadata Standards: Supporting Efficient Data Management in the Field, Laboratory, and Online

Joan Damerow | LBNL

LBNL's Joan Damerow explained that the DOE's Environmental Systems Science Data Infrastructure for a Virtual Ecosystem repository contributors often work in large teams and send physical samples to multiple analysis facilities. This community needs an efficient system for persistent sample identification and tracking that is suitable for the field, laboratory analyses, and online publication. LBNL is conducting a pilot test on the use of persistent identifiers for physical samples. The goal is to provide practical recommendations for efficient sample data management while maximizing the potential value of samples into the future.

11. Metall (Meta Allocator for Persistent Memory) | Keita Iwabuchi (LLNL)

Keita Iwabuchi | LLNL

Data-intensive applications play essential roles across many real-world data analytics domains, often requiring data storage beyond a single process lifetime. LLNL's Keita Iwabuchi introduced Metall, a C/C++ allocator that provides simplified memory allocation interfaces on top of file-backed memory mapping. Metall can maintain allocation management data in compact data structures during operation, increasing locality, and flush the management data to the persistent file during synchronization. Metall has shown up to 3.17x better performance than a similar concept of memory allocator implementation.

12. Finding a Needle in the Haystack: Image Feature Similarity and Feature Extraction at Scale

Travis Thurber | PNNL

As Travis Thurber explained, PNNL extracts image features at scale on AWS to create a compelling image feature similarity solution. The team utilizes AWS Lambda's serverless autoscaling capabilities and Amazon EC2 graphics processing unit instances to process tens of millions of images. Thurber discussed the use of Amazon DynamoDB (for feature storage) and Facebook's dense vector library to make the images queryable through a simple user interface. The poster included advantages and disadvantages of different approaches, lessons learned, and how this approach helps PNNL's sponsors.

13. Modern Infrasound Analysis Tool Suite

Uen-Tao Wang | SNL

A mission-critical need is emerging for visualizing, analyzing, and processing global infrasound events. To classify and understand historic infrasound data, analysts need to perform complex processes and transformations on pre-existing waveform data. Uen-Tao Wang introduced SNL's modern tool set for infrasound analysis—a web-based front-end, a service gateway, and composable, language-agnostic processing and waveform access services—that offers an accessible entry-point for scientists to monitor live data produced by infrasound stations, visualize and analyze historic infrasound data, and interface with algorithm library services.

14. Using the Common Workflow Language for Continuous Post-Processing Pipelines

Sterling Baldwin | LLNL

LLNL's Sterling Baldwin described the Common Workflow Language (CWL)—a workflow specification language for creating reproducible, parallel, composable workflows—in the context of the DOE's Energy Exascale Earth System Model (E3SM) project. Moving E3SM's significant data post-processing needs from bash scripts to CWL workflows has increased the processing rate by reducing idle time. Although CWL has widespread adoption in bioscience and genomics, it has yet to see substantial usage by the climate science community. Baldwin's poster outlines E3SM's current and future CWL work to improve resource efficiency.

15. Nuclear Domain Ontology

Jeren Browning | INL

As Jeren Browning explained, INL's DIAMOND (Data Integration Aggregated Model and Ontology for Nuclear Deployment) framework aggregates the format of the various data sources within a nuclear power plant into one location. This process creates a common structure that can be used for exchanging data between applications. Rather than painful point-to-point integrations, applications can integrate with the ontology to seamlessly communicate with all other applications that have done the same. DIAMOND has been built with extensibility in mind so that modifications can easily be made to accommodate any application.

16. Scientific Image Analysis in Python: From Management to Decisions

Suren Byna | LBNL

LBNL's Suren Byna outlined a joint effort between DOE teams to develop scientific image analysis in Python that provides data pipeline software and enables decisions using experimental data acquired at DOE facilities. Byna highlighted statistical methods to analyze 3D microtomography images and visualize fiber beds, including the introduction of a lossless data reduction algorithm based on maximum projection to detect specimen bulk. This software uses a convolutional neural network to evaluate results from automated fiber detection models, and can help discover and design materials of interest to the DOE.

17. Streaming Deep Learning Image Analysis Pipeline

Gideon Juve | PNNL

Applying deep learning models to streaming data sources can be difficult when the streams vary in size, involve both text and linked images, and require different models to be applied to each image. As Gideon Juve explained, PNNL has developed a deep learning image classification pipeline that addresses these challenges. The pipeline uses a reactive, message-driven, serverless architecture based on AWS Lambda and Amazon Simple Queue Service, while Amazon DynamoDB was used to solve difficult serverless state management challenges.

18. A Scalable and Flexible Data Ingestion Pipeline for Large Multimodal Experiments

Steven Magana-Zook | LLNL

Researchers are being challenged to ingest, fuse, and analyze data from a variety of modalities at scales unfathomable in recent history. LLNL's Steven Magana-Zook described the MINOS (Multi-Informatics for Nuclear Operations Scenarios) project—a data management pipeline with a foundation in Apache Nifi and extended to support file parsing, quality control, and inline analytics. MINOS allows researchers to efficiently perform analytics on the data where it lives in LLNL's cluster. The system also utilizes containers and runs on multiple operating systems. This research is part of a large NA-22 venture.

19. Data Structure for Dynamic Visualizations of Large Time-Series Datasets

Nicole Feist | SNL

Large datasets introduce several challenges for visualizations, and fundamentally, a web browser has limited available memory. To conserve bandwidth and minimize render time, it is desirable to send only the portion of data necessary for visualization at a particular pan and zoom level. Nicole Feist presented SNL's down-sampling strategy for visual representation alongside commonly used techniques. The poster also showed a novel hierarchical data structure for efficient retrieval of tiled time-series data.

20. Tools for HPC File Management

Elsa Gonsiorowski | LLNL

HPC clusters are becoming increasingly complex, partly due to the addition of new storage tiers such as burst buffers. Each tier may have different performance characteristics and may be subject to different policies regarding availability or allocation. Elsa Gonsiorowski summarized tools under active development at LLNL to improve file set management, both between tiers and within a single tier. The poster highlighted the challenges of current HPC storage hierarchies, common HPC data management use cases, and available tools.

21. Data Management Platform for LLNL Life Extension Programs

Sam Eklund | LLNL

Sam Eklund introduced a data management platform for life-extension programs at LLNL. All test and material characterization data are maintained in a central repository with multi-tiered storage, which is indexed to allow searching by test name, material name, or metadata keywords. The team integrates data analysis tools with the repository to allow analysts to work directly with the data on the platform. Additionally, the materials database can automatically generate input decks for HPC code simulations.

Conclusion and Recommendations

Data is a valuable asset for the DOE, whose laboratories and agencies have unique needs, constraints, and resources when it comes to data management. For example, sophisticated HPC systems generate massive amounts of data during simulation runs, while state-of-the-art experimental facilities produce data from disparate sources. As a federally funded research complex, the DOE must make unclassified data available and interpretable by external consumers, including the public. With Big Data opportunities and methodologies quickly outpacing those of other research areas, DOE institutions cannot afford for data management to be merely appended to research programs or project plans. Data, in all its forms and with all of its challenges, deserves a starring role in the DOE's scientific and technological progress.

The breadth and depth of work presented at D3 further illuminated both the importance of data management in these organizations and the innovative solutions DOE teams have developed. To maintain momentum, many participants agreed to establish collaborative spaces for sharing content and continuing discussions—for example, possibly organizing a birds-of-a-feather event at next year's Supercomputing Conference (SC20) and similar upcoming meetings.

Participant feedback, both general and specific, was mainly positive, and the event surpassed the organizers' expectations. An online survey was distributed to all participants, about 21% of whom replied. For more survey data, please see the [Survey Results](#) appendix.

The D3 organizing committee looks forward to next year, and LLNL invites other labs to host the event if desired. For additional coverage, see the news article by [LLNL's Public Affairs Office](#).

Appendices

Organizing Committee

The inaugural D3 event was organized by four LLNL staff with program support.



Jessie Gaylord is a group leader in LLNL's Global Security Computing Applications Division. Her work focuses on multiple aspects of data management including scalable and flexible infrastructures for data ingestion and curation, platforms for data sharing, data engineering, interfaces for data discovery, and data integration. Ms. Gaylord received an M.S. with distinction in computer science from California State University, Chico, and graduated cum laude from Washington University in St. Louis with a bachelor's degree in economics and a minor in mathematics. Previously Ms. Gaylord worked as a business intelligence application developer for NIF and as a market analyst for commercial industry.



Dr. Abdulla holds a Ph.D. (1998) and M.S. (1993) in computer science from Virginia Tech and a bachelor's degree in electrical engineering from Yarmouk University in Jordan. Dr. Abdulla previously worked for the Dow Chemical Company as an IT Specialist. Since joining LLNL in 2000, he has embraced projects that depend on teamwork and data sharing. His tenure includes establishing partnerships with universities seeking LLNL's expertise in HPC and large-scale data analysis. He supported approximate queries over large-scale simulation datasets for the AQSim project and helped design a multi-petabyte database for the Large Synoptic Survey

Telescope. Abdulla used ML to inspect and predict optics damage at the National Ignition Facility, and leveraged data management and analytics to enhance HPC energy efficiency. Recently, he led a Cancer Registry of Norway project developing personalized prevention and treatment strategies through pattern recognition, ML, and time-series statistical analysis of cervical cancer screening data. Today, Abdulla is the principal investigator of the ESGF—an international collaboration that manages a global climate database for 25,000 users on 6 continents.



Dr. Daniel Laney is a computer scientist and group leader at LLNL's Center for Applied Scientific Computing. His research interests include high-performance computing workflow and data management methods, simulated radiographic diagnostics, scientific visualization, and applications of ML to scientific data analysis. Dr. Laney earned a Ph.D. in engineering and applied science at the University of California, Davis, in 2002 and a B.S. in physics from the College of Creative Studies at the University of California, Santa Barbara, in 1996. He joined LLNL in 2002, and currently leads the HPC Workflow project in the WCI directorate.



With a Ph.D. in seismology from Stanford University (1993), Dr. Stan Ruppert is the Geophysical Monitoring Program IT project lead, software team lead, and the LLNL lead for the Low Yield Nuclear Monitoring Dynamic Networks venture. He has been working in a computer science capacity for over 25 years and currently manages the petabyte-scale enterprise IT infrastructure for the Global Security Geophysical Monitoring Program (GMP). Dr. Ruppert provides systems engineering and IT consulting to more than 300 funded programs within LLNL Global Security at several classification levels. He has helped evolve the GMP infrastructure from flatfiles (kilobytes) through database-enabled tools (terabytes), and is supporting the new data-intensive re-architecture to meet current Big Data challenges both at LLNL and with collaborating multi-Lab ventures.

Attendees

Figure 1. Nearly 100 D3 participants gathered outside LLNL's Building 170 for a group photo, taken by Garry McLeod (LLNL).



Attendees represented a very diverse population of people from 21 different organizations and over 20 areas of technical expertise. The following graphics show attendee demographics according to organization, areas of technical expertise, and job titles.

Figure 2. Attendee organization as a percentage of total attendees.

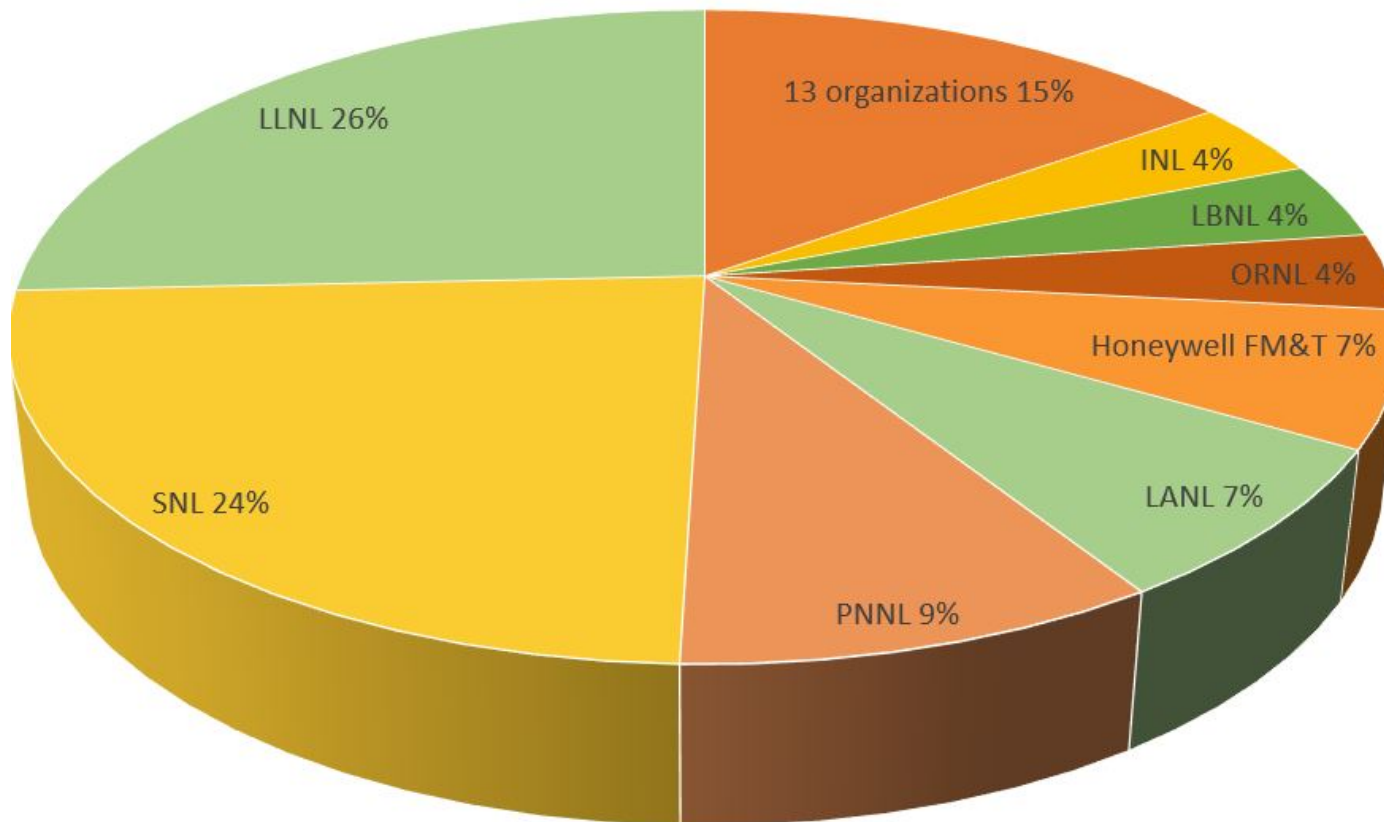
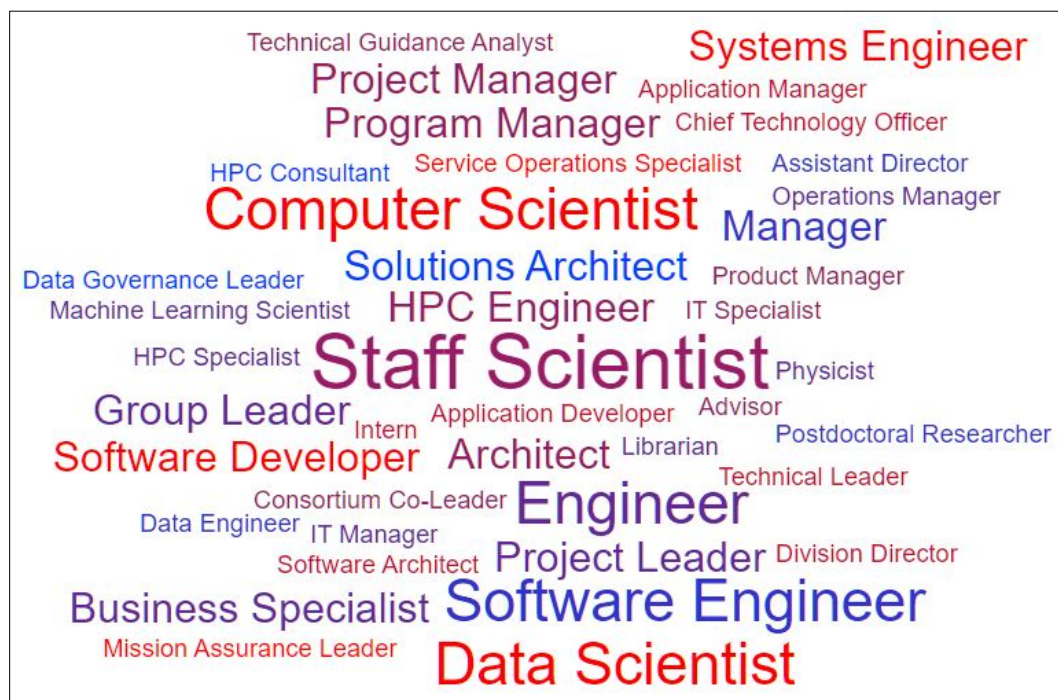


Figure 3. Attendee area of expertise. Larger phrases in the word cloud indicate higher frequency.**Figure 4.** Attendee job title. Larger phrases in the word cloud indicate higher frequency.

The attendee list does not include LLNL support staff.

No.	Attendee Name	Affiliation
1	Ghaleb Abdulla	LLNL
2	Philip Adams	LLNL
3	Sasha Ames	LLNL
4	Kevin Athey	LLNL
5	Sterling Baldwin	LLNL
6	Phillip Baxley	SNL
7	Nick Blazier	SNL
8	Tammie Borders	INL
9	Elizabeth Brown	Honeywell FM&T
10	Jeren Browning	INL
11	Joseph Bruscato	Honeywell FM&T
12	Jeffrey Burke	Kansas City
13	Megan Burns	SNL
14	Suren Byna	LBNL
15	Susan Byrnes	SNL
16	Brian Cain	LANL
17	Allan Casey	LLNL
18	Jacob Cinciripini	GET-NSA
19	Clinton Cohagan	LLNL
20	John Collins	LLNL
21	Aaron Comen	SNL
22	Giovanni Cone	Triad National Security, LLC
23	Joan Damerow	LBNL
24	Jeff Daniels	Honeywell FM&T
25	Kristian Dehaan	Honeywell FM&T

No.	Attendee Name	Affiliation
26	William DeRaad	SNL
27	Charles Doutriaux	LLNL
28	Sam Eklund	LLNL
29	David Emberson	Hewlett Packard Enterprise
30	Nicole Feist	SNL
31	David Fox	LLNL
32	Heath French	SNL
33	Jessie Gaylord	LLNL
34	Lisa Gerhardt	LBNL
35	Dan Goldman	LLNL
36	Elsa Gonsiorowski	LLNL
37	Clay Hagler	PNNL
38	Michael Ham	LANL
39	Craig Hanna	SNL
40	Marcus Hanwell	Kitware
41	Dustin Harvey	LANL
42	David Henderson	Y-12
43	Keita Iwabuchi	LLNL
44	Stephen Jackson	SNL
45	Jennifer Johnson	LLNL
46	Gideon Juve	PNNL
47	Jason Kincl	ORNL
48	Martin Klein	LANL
49	Katie Knight	ORNL
50	Katie Knobbs	PNNL
51	Matthew Kunz	INL

No.	Attendee Name	Affiliation
52	Daniel Laney	LLNL
53	Rebecca Levinson	SNL
54	Rebecca Lewis	NNSA
55	Jennifer Lewis	SNL
56	Matt Macduff	PNNL
57	Paul Madsen	Honeywell FM&T
58	Steven Magana-Zook	LLNL
59	Elaine Martinez	SNL
60	Julie Maze	LANL
61	Kshitij Mehta	ORNL
62	Mark Miller	LLNL
63	James Mitchell	SNL
64	Richard Moleres	SNL
65	Debbie Morford	LLNL
66	Kent Nix	Pantex
67	Ron Oldfield	SNL
68	Gregory Orndorff	SNL
69	Amedeo Perazzo	SLAC
70	Dallin Pew	MSTS
71	Arturo Pino	Honeywell FM&T
72	Paul Pope	LANL
73	Sam Reeve	LLNL
74	Thomas Reichert	SNL
75	David Richards	LLNL
76	Chris Ritter	INL
77	Kelly Rose	NETL

No.	Attendee Name	Affiliation
78	Chad Rowan	NETL
79	Oliver Ruebel	LBNL
80	Stanley Ruppert	LLNL
81	Lee Senter	MSTS
82	Chitra Sivaraman	PNNL
83	Gregory Sjaardema	SNL
84	Eric Stephan	PNNL
85	Theodore Stirm	LLNL
86	Sara Studwell	OSTI
87	Thomas Suckow	PNNL
88	Robert Sutherland	LANL
89	Gary Templet	SNL
90	Matthew Templeton	Honeywell FM&T
91	Sandy Thompson	PNNL
92	Travis Thurber	PNNL
93	Malachi Tolman	SNL
94	Ron Trujillo	LANL
95	Greg Tubbs	LLNL
96	Craig Ulmer	SNL
97	Otto Venezuela	LLNL
98	Rick Wagner	ANL
99	Lipeng Wan	ORNL
100	Uen-Tao Wang	SNL
101	Mary Beth West	OSTI
102	Lisa Wilkening	SNL
103	Lynn Wood	PNNL

No.	Attendee Name	Affiliation
104	Brian Young	SNL
105	Chengzhu Zhang	LLNL

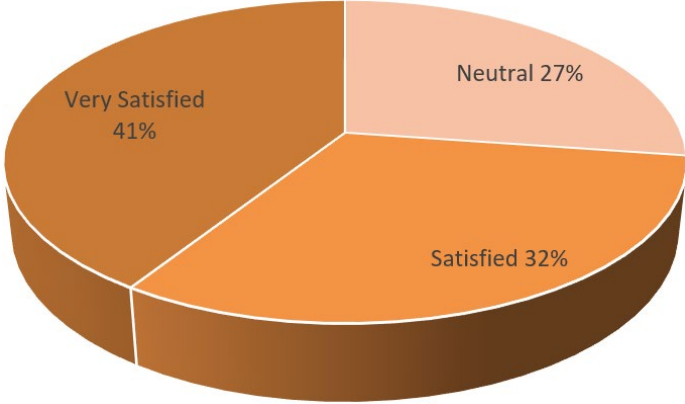
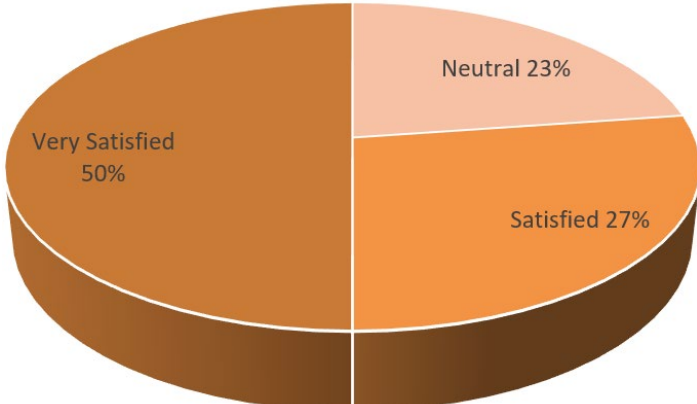
Survey Results

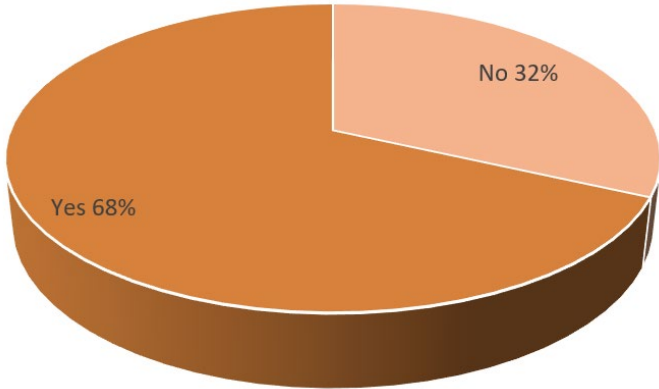
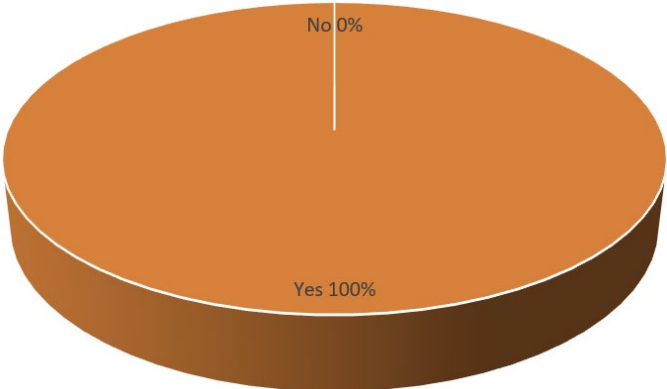
Of the 105 event participants, 22 completed the online survey for a 21% response rate. D3 organizers and support staff excluded themselves from the survey. Many questions asked respondents to rate various aspects of D3 on a five-point scale (Very Satisfied, Satisfied, Neutral, Dissatisfied, Very Dissatisfied), while some questions were in Yes/No format. Respondents were able to enter free-form comments throughout. D3 organizers appreciated all respondents' thoughtful feedback.

The survey drew a meaningful sample of attendees from the invited organizations as well as a range of technical expertise and interests. Most respondents appreciated hearing about data management strategies at other DOE organizations and welcomed face-to-face interactions. The overall sentiment shared by most respondents was that D3 was a valuable event with a necessary future.

The results highlight several opportunities for improvement. For example, while many respondents heard about D3 through colleagues or via email, others learned of the event relatively late or by chance. Facility tours were praised, though some respondents noted the main venue's shortcomings for a workshop of this size (e.g., the food buffet was set up in a hallway). In addition, respondents suggested many topics for inclusion in the next event's agenda, such as data anomaly detection, data catalogs, and bioinformatics.

Responses (n = 22)	Sample Comments	Improvement Opportunities								
Overall, how would you rate this event?										
<table><tr><th>Rating</th><th>Percentage</th></tr><tr><td>Satisfied</td><td>50%</td></tr><tr><td>Very Satisfied</td><td>41%</td></tr><tr><td>Neutral</td><td>9%</td></tr></table>	Rating	Percentage	Satisfied	50%	Very Satisfied	41%	Neutral	9%	<ul style="list-style-type: none">• “Long overdue, well attended, very diverse array of work and perspectives, working groups avoided death by PowerPoint.”• “We didn’t come from a national lab so only heard about the event at the last minute by word of mouth.”• “It was great to hear about the work other teams are doing. I picked up a lot of useful information.”	<ul style="list-style-type: none">• Use survey feedback to inform and prioritize future D3 agendas.• Promote the event earlier and more widely.• Find a larger venue with rooms/areas better suited to different aspects of the event (e.g., dedicated area for posters instead of around the perimeter of the main room).
Rating	Percentage									
Satisfied	50%									
Very Satisfied	41%									
Neutral	9%									

Responses (n = 22)	Sample Comments	Improvement Opportunities								
Based on your attendance, rate the range of topics presented during D3										
 <table border="1"><thead><tr><th>Satisfaction Level</th><th>Percentage</th></tr></thead><tbody><tr><td>Very Satisfied</td><td>41%</td></tr><tr><td>Satisfied</td><td>32%</td></tr><tr><td>Neutral</td><td>27%</td></tr></tbody></table>	Satisfaction Level	Percentage	Very Satisfied	41%	Satisfied	32%	Neutral	27%	<ul style="list-style-type: none">• “A narrower scope would have been more appropriate. I would separate out the data meeting from ‘models’ meeting.”• “Although I'd hate to miss any of the presentations, the range of topics was so broad that it might be more effective to have concurrent tracks rather than everyone presenting in one room.”• “It would be good to hear from application end-users about their data management issues.”	<ul style="list-style-type: none">• Specify scope in more detail in call for abstracts.
Satisfaction Level	Percentage									
Very Satisfied	41%									
Satisfied	32%									
Neutral	27%									
Based on your attendance, rate the quality of topics presented during D3.										
 <table border="1"><thead><tr><th>Satisfaction Level</th><th>Percentage</th></tr></thead><tbody><tr><td>Very Satisfied</td><td>50%</td></tr><tr><td>Satisfied</td><td>27%</td></tr><tr><td>Neutral</td><td>23%</td></tr></tbody></table>	Satisfaction Level	Percentage	Very Satisfied	50%	Satisfied	27%	Neutral	23%	<ul style="list-style-type: none">• “More specific analytics topics and updates to some of the data management solutions that were presented.”• “More information from users of systems that are being built around the DOE.”• “We should continue but set the bar higher for the quality of presentations.”	<ul style="list-style-type: none">• Consider including classified subject areas (with appropriate security procedures and venue).• Broaden the abstract review committee to include more people and possibly other labs
Satisfaction Level	Percentage									
Very Satisfied	50%									
Satisfied	27%									
Neutral	23%									

Responses (n = 22)	Sample Comments	Improvement Opportunities
As a result of D3, do you plan to collaborate with other sites?		
 <p>Yes 68%</p> <p>No 32%</p>	<ul style="list-style-type: none"> • “We were already collaborating with some attendees, but we also made a new connection.” • “There was talk of maybe creating a way for folks working in similar fields to be in touch. I would be very much interested in this.” • “I plan to collaborate with many sites on data governance and data science architecture and pipelines.” 	<ul style="list-style-type: none"> • Encourage collaborations outside of D3 via the D3 email list, the D3 website, an online collaboration space, or through meet-ups at related events.
Do you think there should be another D3?		
 <p>No 0%</p> <p>Yes 100%</p>	<ul style="list-style-type: none"> • “The logistics went off very well. It was obvious the LLNL team are experts in Data Management and facilitation. Looking forward to next year (or sooner)!!” • “The most valuable thing was meeting and talking to other DOE people who work in this area.” • “I like that it was open to non-DOE employees for larger inclusion. In some cases this would preclude some projects or sites from being shared.” 	<ul style="list-style-type: none"> • Consider inviting other organizations that work with DOE labs and agencies. • Balance agenda so facility tours, working groups, and poster sessions have less competition for attendees’ time. • Schedule the next workshop farther away from fiscal year end.

Acronyms

Acronym	Definition
ACTICI	Advanced Computer Tools to Identify Classified Information program
ANL	Argonne National Laboratory
API	Application programming interface
ASCR	Advanced Scientific Computing Research program
AWS	Amazon Web Services
CI	Continuous integration
CWL	Common Workflow Language
D3	DOE Data Day
DIAMOND	Data Integration Aggregated Model and Ontology for Nuclear Deployment
DNN	Defense Nuclear Nonproliferation
DOE	U.S. Department of Energy
DOI	Digital object identifier
E3SM	Energy Exascale Earth System Model
EDX	Energy Data eXchange
eOS	Ecosystem for Open Science
ESGF	Earth System Grid Federation
FAIR	Findable, accessible, interoperable, reusable
FM&T	Federal Manufacturing & Technologies (Honeywell)
GMP	Geophysical Monitoring Program
HDMF	Hierarchical data modeling framework
HPC	High-performance computing
INL	Idaho National Laboratory
IT	Information technology
LANL	Los Alamos National Laboratory
LBNL	Lawrence Berkeley National Laboratory
LCLS	Linac Coherent Light Source

Acronym	Definition
LLNL	Lawrence Livermore National Laboratory
MDM	Master data management
MINOS	Multi-Informatics for Nuclear Operations Scenarios
ML	Machine learning
MSTS	Mission Support and Test Services
NA-22	Nonproliferation Research and Development program at NNSA
NETL	National Energy Technology Laboratory
NIF	National Ignition Facility
NNSA	National Nuclear Security Administration
NSE	Nuclear Security Enterprise
NTK	Need to know
NWB:N	Neurodata Without Borders: Neurophysiology
ORNL	Oak Ridge National Laboratory
OSTI	DOE Office of Scientific and Technical Information
PDC	Proactive Data Containers
PNNL	Pacific Northwest National Laboratory
PRIDE	Product Realization Integrated Digital Enterprise
REST	Representational state transfer
SEDS	Stockpile Evaluation Data System
SLAC	Stanford Linear Accelerator National Laboratory
SNL	Sandia National Laboratories
WCI	Weapons and Complex Integration
Y-12	DOE National Security Complex