DOE Data Day 2020 Report

October 5–7, 2020
Virtual Conference

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

Organizing Committee

Tammie Borders, Angela Sheffield, Marc Wonders (NNSA HQ)
Shiloh Elliott, M. Ross Kunz, Christopher Ritter (INL)
Jeffrey Burke, Ruben Pino (KCNSC)
Martin Klein (LANL)
Ghaleb Abdulla, Loni Cason, Jessie Gaylord, Dan Laney, Angeline Lee, Stanley Ruppert (LLNL)
Kelly Rose (NETL)
Gideon Juve, Sandra Thompson (PNNL)
Katherine Anderson Aur (SNL)

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-AR-819828

.

Contents

Executive Summary

The Department of Energy's (DOE) continued engagement with the larger scientific community in the promotion of data management led organizing the Second Annual DOE Data Days (D3) workshop. D3 serves as a means to higher-quality and more efficient research and analysis, in addition to serving as a critical component of data science within DOE programs.

The second annual D3 workshop was held on October 5-7, 2020, in a virtual (remote presence) manner due to COVID-19 travel restrictions and protocols, convened by Lawrence Livermore National Laboratory (LLNL). It was organized by a multi-laboratory committee in an effort to bring data management practitioners at the DOE laboratories together to share their work and results, facilitating knowledge transfers and best practices across project teams. Tools and platforms to support data management and analysis are rapidly evolving and provide enormous opportunities. The workshop featured 37 Data Science (DS) Portfolio researchers as presenters, panelists, and session moderators, bringing together 164 attendees from 28 organizations, including DOE laboratories, National Nuclear Security Administration (NNSA) headquarters, and a handful of university and mission partner organizations. Initial survey results reflected success despite the virtual format this year. Twenty-three presentations were grouped into four themed sessions, each with a plenary talk, followed by session topic areas focused on a wide range of challenges that can be specific to DOE but are common across DOE mission areas and organizations. A fifth non-themed three presentation session was added to address specific use-case data scenarios. This year's workshop also launched the concept of D3-focused hackathons specific to data management issues within DOE.

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 laboratories. The response was overwhelmingly supportive and 74 abstract submissions were received. Themes emerged from the abstract submissions, so sessions were organized into four topic areas:

- Data Curation and Standards: Legacy Data, Existing Data, and Future Data
- Data-Intensive Computing, High Performance Computing (HPC), and Tools for DOE's Computing Communities
- Cloud, HPC, and Hybrid Data Management
- Data Access, Sharing, and Sensitivity

These subject areas provided the framework for the agenda, which featured virtual talks given by Argonne National Laboratory (ANL), Brookhaven National Laboratory (BNL), DOE Office of Scientific and Technical Information (OSTI), Idaho National Laboratory (INL), Joint Genome Institute (JGI), Lawrence Berkeley National Laboratory (LBNL), LLNL, Los Alamos National Laboratory (LANL), National Energy Technology Laboratory (NETL), NNSA, National Renewable Energy Laboratory (NREL), Pacific Northwest National Laboratory (PNNL), Sandia National Laboratories (SNL), and University of Tennessee, Knoxville (UT Knoxville) researchers. These topics were also the basis for a moderated panel discussion and question and answer (Q&A) period following each distinct session. A fifth session was added to discuss highlighted Data Curation Collection Specific Use Cases.

Participant discussions and engagement during the workshop were phenomenal. There was a clear consensus that the workshop was successful even though held virtually, with fewer accepted talks possible due to the shortened half-day virtual format. In addition, motivation and concept for a D3 organized technical demonstration exercise ("Hackathon") was introduced and initial design and planning instantiated.

This Report summarizes the important discussions and recommendations from the different working sessions and contains the agenda, submitted abstracts, virtual talks, 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, abstracts, 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 Principal Directorate (WCI), and the efforts of the D3 multi-laboratory organizing committee members for abstract review and distillation into theme areas, as well as topic-area session chairs and panel moderators. Event logistics, planning, and virtual format tools (CVENT and Zoom) were led by Loni (Hoellwarth) Cason (LLNL), who successfully navigated the many challenges in converting the D3 workshop to a virtual format with rapidly evolving tools and capabilities.

In particular, NA-22 Data Science Program Manager Angela Sheffield provided funding for planning, abstract reviews, and this Report. WCI provided access to administrative support staff.

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 first annual D3 workshop was born from this critical work and held on September 25–26, 2019. The second annual D3 workshop, held October 5–7, 2020, continued the collaborative discussions through a series of three half-day technical sessions comprised of curated talks followed by a moderated discussion and Q&A session to foster dialogue and engagement in this year's virtual setting. The panel discussions encouraged inter-participant conversations. The 2020 D3 event received 215 registrations and welcomed than 164 participants from across the DOE complex—almost double the 2019 attendance.

D3's continued 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 28 presentations and moderated panels the event was deemed a success. Participant feedback indicated a strong preference for continuing the annual D3 workshops as well as support for additional platforms such as hackathons to gather early-career staff together to work collaboratively on technical challenges. The D3 workshops continue to fill 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.

Data Management Challenges

Most programs at the national laboratories 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. However, often these efforts exist in programmatic stovepipes. The goals remain 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.

The Need for DOE Data Days

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 the broader science questions of tomorrow.

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 the National Science Foundation (NSF) and National Aeronautics and Space Administration (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 laboratories. 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 laboratory-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 laboratories to discuss the particular challenges of developing standards-based processes and systems to manage volumes of national security data in laboratory environments. Since data management is a support function for other work, cross-program and cross-laboratory 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.

DOE Data Days (D3)

A recurring (annual) workshop dedicated to data management work at the DOE national laboratories (named DOE Data Days, or “D3” for short) provides an extremely valuable forum for data management practitioners and system developers. Many programs are investing more formally in data management, and open discussions are critical to make efficient progress in this fast-moving field, promoting shared solutions and best practices that are effective in laboratory environments. Presentations and discussions on data, software, storage, and network topics specific to laboratory programs and constraints have proven enormously valuable to multiple missions. Topics have included (but are not limited to):

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

- Commercial cloud usage at the labs
- Convergence of high-performance computing, big data, and cloud
- Data sharing, processing, and archiving across isolated and classified networks and facilities
- Moving, managing, and storing large volumes of data
- Multi-laboratory authentication, cyber approvals, and other data-security considerations
- Data archiving, processing, and sharing on classified networks
- Curating experimental and large-scale simulation data

Developers, data managers, data generators (including scientists/engineers/analysts), researchers, and information technology (IT) support personnel at the national laboratories have been encouraged to participate in this event. Presentations have highlighted developing approaches and effective existing solutions in a variety of scientific domains. Informal or organized discussions have facilitated information sharing, collaborations, and better integrations between programs. The objective of the ongoing workshop series is to continue promoting awareness of effective data management strategies, shorten the learning curve for new efforts, and increase the overall quality of data management practices at the national laboratories.

Repeat events are currently planned on an annual basis with topic areas evolving to support DOE's NA-22 mission priority and areas of interest.

Agenda and Abstracts

The virtual D3 workshop was organized into four themed sessions: Data Curation and Standards – Legacy Data, Existing Data, and Future Data; Data Intensive Computing, HPC, and Tools for DOE's Computing Communities; Cloud, HPC, and Hybrid Data Management; and Data Access, Sharing, and Sensitivity. A fifth session addressed specific data use cases. The D3 event kicked off with opening and keynote presentations. The "NNSA Headquarters Perspective and Value-Goals of D3" opener was given first by Tammie Borders (NNSA Headquarters), followed by a keynote, "Data Science Machinations, Musings and Eureka Moments at the Joint Genomic Institute (JGI)," given by Kjiersten Fagnan. Following the presentations, each of the four theme sessions included a plenary talk, which introduced the session topics, followed by individual presentations. The sessions concluded with a moderated panel discussion and Q&A period from D3 attendees via an online question submission tool. The plenary speakers were each allotted 20 minutes and session speakers had 15 minutes for their presentations. Some session talks were pre-recorded video streams, others were given live via Zoom. All talks were recorded, as were the moderated panel discussions and Q&A. All speakers were present during the moderated panel discussions and Q&A. During the moderated panel discussions, the moderator facilitated discussions covering their session talks and any submitted questions. The 2020 D3 workshop was summarized with concluding thoughts and next priorities given by Angela Sheffield (NNSA Headquarters). See the event website proceedings tab for full abstracts, and presentation slides. While sessions were recorded, for security reasons, it has proven impractical to make the recordings widely available. The summaries below are reflective of the presentation abstracts provided.

Session 1: Data Curation and Standards – Legacy Data, Existing Data, and Future Data

The first D3 session featured speakers on topics related to the provenance and classification of data. Attendees learned about data standards documentation, metadata schemas, data quality ratings, databases and tools, the reproduction and tracing of scientific workflows, updating legacy data formats, and the assessment of metadata.

Streamlining Data Standard Documentation Through GitHub Integration

Robert Crystal-Ornelas | LBNL

Data that are submitted to repositories and adhere to data standards are more interoperable and easily used in data integration, said Robert Crystal-Ornelas of LBNL in his plenary talk. The U.S. DOE's Environmental Systems Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE) repository worked with six teams of community partners from the network of national laboratories to develop 7 metadata/data standards related to their research domain. He shared how the community partners adopted GitHub's version control system to solve the dual issues of keeping standards both responsive to contributor suggestions and also user-friendly. All community partners drafted their data standards in CSV files. They created CSV parsers to automate conversion into markdown code for upload to GitHub. Because data standards are on GitHub, contributors can easily submit a GitHub "issue" that is sent directly to the community partners that created the standard. GitHub issues are visible to the public and help prioritize updates to the standard. The final step in the data standard publication process takes advantage of GitHub's integration with a documentation-

building system called gitbooks. In gitbooks, markdown code on GitHub is rendered into an easy-to-use website, which can be read and shared by those not familiar with the GitHub interface.

DCAT-eOS-AP: A Metadata Schema for Use in Nuclear Monitoring Projects with Applications to Other DOE Mission Areas

Katherine Anderson Aur | SNL

Katherine Anderson Aur of SNL explained that metadata are instrumental in the interpretation and effective use of data, particularly by those who are not the primary creator. Accurate metadata facilitates data discovery, as well as the understanding of the purpose of a project or dataset and can aid in interpretation and analysis during initial use and reuse by future researchers. Since they describe data structure and standards, metadata schemas are thus important for dataset discovery, use, reuse, preservation, and sharing. There has not previously been an effort to develop a metadata schema that could be applied across a broad range of projects within the DNN Research and Development (R&D) portfolio. Increasing the standardization of metadata across projects will increase the value of the data at the program level and throughout the DNN R&D portfolio. They developed a machine-readable metadata schema for several domains relevant to nuclear monitoring. This new schema, DCAT-eOS-AP, builds on existing ontologies, to the extent possible, to promote interoperability and discovery. Due to the multitude of phenomenologies represented in this schema, it is anticipated that it will be easily extensible to various projects across many DOE mission areas.

Development of Data Quality Ratings for Experimental Alloy Data Through DOE's eXtremeMAT Consortium

Madison Wenzlick | NETL

The ability to collect, curate and analyze large volumes of data has enabled the development of data analytics tools for a wide array of applications, including in the field of materials design, said NETL's Madison Wenzlick. In particular, the design of alloys focuses on understanding the relationships between the material composition, processing and heat treatment, and the mechanical properties. There is a wide variation in data reporting, including in metadata, standards of experimentation, data format, availability and trustworthiness of the source itself. Therefore, the data must be segmented according to its quality in order to establish the validity of the proceeding analytics. Through the U.S. DOE's eXtremeMAT Consortium, a joint PNNL-NETL team has focused on collecting and curating data from open-source literature, as well as extracting data from online databases and datasheets to create a database on alloy properties. In order to address possible discrepancies in the quality of data collected, the eXtremeMAT team developed metrics for rating the quality of data and information on experimentally obtained alloy properties, processing and composition. eXtremeMAT determined ratings of quality from 1-5 according to the data completeness, accuracy, usability, and standardization. These ratings have been applied to datasets pertinent to the eXtremeMAT project, allowing for higher quality data to be used in predictive analytics, and lower quality data to be used in validation. A rating system has been implemented within DOE Fossil Energy's Energy Data eXchange (EDX), which enables information exchange among researchers across DOE laboratories, facilitates the determination of data quality from crowd-sourced ratings, and automates the addition of data quality rating to the source metadata.

High Throughput Experimental Materials Database and Related Data Tools for the Experimental Laboratory

Kevin Talley | NREL

The materials science experimental and computational modeling communities require accessibility to stores of experimental synthesis and characterization data and metadata, said Kevin Talley of NREL. The High Throughput Experimental Materials (HTEM) Database (http://htem.nrel.gov) provides an important resource of materials data by combining modern data management with experimental materials science workflows. This more complete data picture enables improved property predictions by giving ML models a direct connection to synthesis conditions and processing variables. He discussed the details of the underlying machine network, extraction protocols, and meta-data collection from the perspective of implementation, adaptability, and end-user experience for all aspects of the data work-flow cycle. This data toolset enables the efficient execution of low-error experimental research and represents the type of databases that applied ML studies require for improved property prediction. In total, these tools guide data through the full experimental cycle, from experiment to publication. HTEM is connected to other external public data resources, sourcing data to Citrine Informatics, several DOE Energy Materials Network (EMN) datahubs, and National Institute of Standards and Technology (NIST) High-Throughput Experimental Materials Repository and Registry. The impact of this work is a standardized and flexible architecture for harvesting, sorting, extracting, and storing valuable experimental data. This is a unique asset because it is a large and chemically diverse public database of experimentally measured quantities that can be accessed by an application programming interface for integration with other digital projects and partnerships. This data toolset accelerates the research cycle and improves standardization of experimental data.

Containerized Environment for Reproducibility and Traceability of Scientific Workflows

Paula Olaya | UT Knoxville

Scientists use simulations to study natural phenomena, said Paula Olaya of UT Knoxville, and trusting the simulation results is vital to the integrity of scientific discovery. To trust results, she said the simulations' reproducibility, replicability, and traceability must be ensured through the annotation of simulation's executions. She and her colleagues proposed an operative system-level solution that leverages the intrinsic characteristics of containers (i.e., portability, isolation, encapsulation, and unique identifiers) to annotate workflows and capture their metadata. The solution enables transparent and automatic metadata collection and access, easy-to-read record trail, and tight connections between data and metadata. They built a prototype of a containerized environment which encapsulates each component of a scientific workflow (i.e., data and applications) in individual containers, features zero-copy data transfer between containers, requires no modification of the underlying applications, and automatically links the metadata to the workflow. They assessed the effectiveness of the prototype for four increasingly complex workflows, ranging from simple visualization applications such as, gnuplot to machine learning applications, in particular weighted k-Nearest Neighbors (kKNN) and random forest (RF); and show their ability to build workflow record trails at the OS-level for all four scenarios in an automatic, easy-to-read, and with a tight connection between data and metadata. The containerized environment addresses metadata from OS system-level by leveraging cutting-edge container technology to provide a complete, transparent and automatic collection and management of workflow metadata.

Modernization of the Evaluated Nuclear Structure Data File (ENSDF)
Elizabeth McCutchan | BNL

The Evaluated Nuclear Structure Data File, ENSDF, contains recommended nuclear structure and decay data for all nuclides that have been observed experimentally, explained Elizabeth McCutchan of BNL. Data from nuclear reactions and decay processes are individually presented along with recommended values for level and gamma-ray properties derived from a critical analysis of all measurements. ENSDF is the definitive, world-wide resource for nuclear structure and decay data. It forms a cornerstone of low-energy academic nuclear physics and astrophysics while providing vital input in various industrial, medical, and homeland security applications. The format used to store ENSDF evaluations as well as the suite of codes required to perform the evaluation work has been developed over many decades. Data in ENSDF are stored in an 80-column mixed-record format, with character flags and column positions indicating the data type. Additionally, comment records are frequently used to denote additional data, provide references and detail the steps taken to arrive at the recommended value. This legacy format obviously makes parsing and manipulating the data in ENSDF extremely challenging. The presentation described the ENSDF database and provided examples on how it has made significant contributions to both fundamental and applied sciences. They then described a new initiative to make the treasure trove of key data within ENSDF more accessible to modern computational tools and search engines. This involves encapsulating ENSDF in an object-oriented database. developing machine learning techniques to streamline the pathway from publication to data evaluation and expanding the database to include observables obtained with modern experimental technologies and applicable to broader community needs.

DOE OSTI's Approaches to Artificial Intelligence and Machine Learning for R&D Results
Mary Beth West and Joshua Nelson | DOE OSTI

DOE OSTI established its artificial intelligence (AI) team in the summer of 2019. As presented by Mary Beth West, the AI Team's work and research in this space are new endeavors for OSTI; identifying the appropriate areas of research and investigation are priorities for the team and will ensure results and products that support OSTI and the collection, preservation, and dissemination of R&D results. To support OSTI's strategic plan, the AI Team has started an assessment of the current R&D results corpus (e.g., metadata and full text) collected through ingest products such as E-Link and DOE CODE and disseminated through OSTI.GOV and other discovery applications. This presentation presented applied AI and ML approaches to assess and address data challenges and discussed how these data challenges are being evaluated to establish a comprehensive corpus of R&D results, support the reuse of R&D results and its data, and extend these findings to the broader DOE community.

Session 2: Data Intensive Computing, High Performance Computing (HPC), and Tools for DOE's Computing Communities

The second D3 session showcased how practitioners can make data more useful and accessible. Data automation, cross-facility scientific collaboration, analysis and management of multidimensional data, large-scale data analysis in the cloud, mining scientific literature, and the intelligent backfill of missing data were topics of discussion.

Gladier: An Architecture to Enable Modular Automation of Data Capture, Storage, and Analysis at Experimental Facilities

Ben Blaiszik | ANL

In this plenary talk, Ben Blaiszik of ANL described Gladier (the Globus Architecture for Data-Intensive Experimental Research), a data architecture that enables the rapid development of customized automation flows linking many data services to enable data capture, storage, sharing, publication, and analysis solutions for experimental facilities. Gladier builds on a collection of data services built for science, including Globus Auth, Transfer, Search, Groups, and Automate, the Materials Data Facility for data publication and discovery, DLHub for model publication, and funcX to enable distributed function-as-a-service execution. Gladier relies on lightweight data management and function execution agents deployed on participating edge systems (e.g., acquisition machines, HPC clusters, storage systems, and even laptops). Deploying Gladier for new use cases is as simple as ensuring these agents are deployed on participating resources and configuring automated flows and access permissions through representational state transfer (REST) application programming interfaces (APIs), Python SDKs, command line interfaces, and web interfaces. Globus services are highly reliable, professionally operated cloud-hosted services that support the work of over 150,000 researchers worldwide as foundational capabilities for scientific applications and workflows; using them greatly reduces the burden on local systems, administrators, and programmers.

The presentation detailed a new deployment of Gladier at Argonne's Advanced Photon Source (APS) coupled with HPC resources at the Leadership Computing Facility (ALCF) across three beamlines - tomography (2-BM), serial synchrotron crystallography (SSX) beamline, and X-ray Photon Correlation Spectroscopy (XPCS 19-ID and 8-ID). At the SSX beamline specifically, automated capture, analysis, transfer, indexing, storage, and cataloging of data from experiments have enabled on-demand analysis and solving of the crystal structure of COVID-19 surface proteins. This and other data are then cataloged and shared with remote collaborators through a custom portal. With each of these beamlines, the team explored new topics in automation and built new modular components for reuse in the future. In the next year, work will continue to harden the current deployment, continue to build new components, and integrate these capabilities at new beamlines.

Cross-Facility Science: The Superfacility Model at Lawrence Berkeley National Laboratory (LBNL)

Deborah Bard | LBNL

As data sets from DOE user facilities grow in both size and complexity there is an urgent need for new capabilities to transfer, reduce, analyze, store, search and curate the data in order to facilitate scientific discovery. As explained by Deborah Bard of LBNL, the National Energy Research Scientific Computing Center (NERSC) has expanded services and designed new capabilities in support of experimental workflows via Advanced Scientific Computing Research's (ASCR's) powerful computing, storage and networking resources. In this talk, the Superfacility concept was introduced - a framework for integrating experimental and observational research instruments with computational and data facilities at NERSC and ESnet. The science requirements that are driving this work, and how this has translated into technical innovations in data management, scheduling, networking and automation, were discussed. The impact of this work using examples of teams that are using our systems for real-time experimental data analysis (such as the LZ dark matter detection experiment, the Dark Energy Spectroscopic Instrument [DESI], and lightsource facilities such as ALS and LCLS), pushing our infrastructure in new ways, were illustrated. In particular, focus was on the new ways experimental scientists are accessing HPC facilities, and the implications for future system design.

Combined Nonnegative Tensor Factorization & Bayesian Nonparametric Models Analyzing and Managing Multidimensional Biota Data

Helen Cui | LANL

Identifying, applying and developing efficient analytic tools are essential for interpreting large, multi-dimensional datasets, especially emerging non-traditional datasets for proliferation detection applications, said Helen Cui of LANL. For decades, physical and chemical signals have been measured, analyzed and studied extensively. Nuclear facilities and activities also impact biological systems in the surrounding environment, potentially resulting in measurable biosignatures. Processing and analyzing multi-dimensional biota data correlating with metadata and other measurements have become a bottleneck to understanding biosignatures. The LANL team is approaching this problem with combined methods of Non-negative Tensor Factorization and Bayesian nonparametric modeling. High-dimensional data are naturally organized in tensors (i.e., multi-dimensional arrays) and tensor factorization is a cutting-edge approach for factor analysis. Its main objective is to decompose the data into factor matrices that carry the latent features in each tensor dimension. LANL's unsupervised learning tool, NTFk, based on Non-negative Tensor Factorization, extracts multimodal and easy interpretable features that expose different manifestations of dominant latent processes buried in the data and enable their identification and characterization. NTFk characterizes the relations among different tensor dimensions, reconstructs missing data in the datasets, and is scalable specifically for discovery-based unsupervised machine learning and big data analytics. The information driven data analyses will be developed to explore multi-omics data to discover the biosignatures that are most informative contributing to multi-phenomenological signatures. Approximate dynamic programming will be applied to sensor selection, using information-theoretic optimization criterion resulting in practical and scalable algorithms for measurement selection in distributed inference problems over long time horizons.

Interactive Large-Scale Seismic Noise Analysis in the Cloud Using Python and Kubernetes

Jonathan MacCarthy | LANL

As research in seismology continues to identify valuable new signals in ever-growing data streams, said Jonathan MacCarthy of LANL, it becomes more important to explore research tools and platforms that can scale from small exploratory analyses to large survey-style applications. The commercial cloud offers a diverse and powerful platform to quickly perform large-scale research, but it also comes with a number of practical challenges. Most existing research software in seismology is not compatible with a remote distributed system like the cloud. Additionally, there is a significant learning curve in using a new software ecosystem. Finally, standard seismic formats, such as miniSEED, SAC or PH5 may not be optimal for access on distributed systems, where the balance between compression, file size and network communication is different compared to local or HPC systems. In this work, the Xarray, Dask and Zarr libraries in the Python software ecosystem were used to address some of the challenges outlined above. Regional and continent-scale seismic noise analysis were performed using the Amazon Web Services (AWS) cloud platform to demonstrate an interactive and fully in-cloud research workflow that accelerates time-to-result.

Scientific Literature Mining for X-ray Absorption Spectroscopy (XAS)

Gilchan Park | BNL

Text mining is the process of automatically extracting meaningful information from large volumes of unstructured text data, information that can be directly presented to users or put into structured formats for populating databases, said Gilchan Park of BNL. Users from academia and industry bring their samples to National Synchrotron Light Source II (NSLS-II) at BNL for characterization of chemical bonding and electron energy band structure with the guidance of beamline scientists. During the short time users spend at the beamline for their experiments (typically several 4-hour sessions over 48 hours), they compare spectrum results of their samples to those of well-characterized reference samples. NSLS-II users have complex information needs that cannot easily be answered by popular search engines such as Google Scholar and Web of Science, and finding comparable spectra in the vast literature during users' time at the beamline is inefficient and haphazard. Park presented a pilot system for scientific literature mining for answering NSLS-II users' complex information needs. The system extracts and presents figures, captions and text related to a specific XAS spectrum from the scientific literature. Users can find spectra using a classification of papers by transition metals and XAS edges or using search on a text collection. The backend models have been built using deep learning based-contextual word representations and domain-specific text mining tools, such as ChemDataExtractor. Park asked users to help with evaluation of relevance using a quick rating system to improve model performance.

Data Challenges in High Energy Physics Workflows

Nathan Tallent | PNNL

High energy physics (HEP) workflows face a myriad of data challenges, said Nathan Tallent and Noah Oblath of PNNL. They discussed two HEP workflows; first, they motivated the need for an intelligent data movement framework for Belle II computing's Monte Carlo Simulations; then, emerging design considerations were discussed for real-time event reconstruction in Project 8. The work is funded by DOE ASCR ("Integrated End-to-End Performance Prediction and Diagnosis") and the U.S.-Japan Science and Technology (S&T) Cooperation Program. The Belle II experiment, searching for New Physics, seeks discrepancies in the predictions of precision measurements of B-mesons rare decays. Although the experiment's instrument is

located at the KEK particle accelerator in Japan, many physicists around the world access its data. To reduce the long access latencies of remote data, the researchers developed the TAZeR (Transparent Asynchronous Zero-copy Remote I/O) data movement framework. Results show that TAZeR can increase workload throughput by several multiples – up to 20× in some cases – and with a data-request rate of 24× the incoming WAN links.

The Project 8 collaboration is pursuing a measurement of the absolute neutrino mass using tritium beta-decay. The unique setup of the experiment requires a data acquisition (DAQ) system that can perform real-time digital beamforming on ~100 antenna channels. Due to the particular characteristics of the beamforming process, the trigger and tracking stages require all of the data (full frequency range) from all channels for a given period of time. This process cannot be parallelized by antenna channel or even beamformed voxel. There are currently a few different algorithms under consideration for performing the triggering and track-reconstruction analysis. The algorithms that are eventually selected will need to operate in real time and have a well-understood efficiency.

Exploring the Use of Machine Learning Algorithms to Backfill Missing Traffic Data

Ambarish Nag | NREL

Traffic engineering relies on information about travel times, vehicle counts, and speeds to optimize flow along a corridor, i.e., a series of intersections between which traffic light timings are synchronized. Traditionally, cities use roadside sensors to capture such information, presented Ambarish Nag of NREL. However, the price for these devices is too high to deploy them at scale. The team is building a digital twin for the traffic system at regional scale for Chattanooga, Tennessee. As part of this project, traffic signal control was optimized along Shallowford Road corridor with the goal of reducing energy use by 18-20%, which we achieve by minimizing braking (a major contributor to energy use). Corridor performance metrics were tracked such as speed (mph) and travel time to traverse the corridor (seconds) at 15-minute intervals once a corridor is defined. The first set of field experiments took place in February 2020, before corridors were defined, which makes it difficult to compare this first set of experiments with later experiments. To improve this comparison, the gap was backfilled from January 3rd to March 16th with high-fidelity estimates. These approaches were compared with a MultipleOutputRegressor (scikit-learn) which uses separate/independent models for each output. Lastly, a RegressorChain of models was created, in which the first model uses the original data as input, and each subsequent model used the output of the previous model as input for its prediction. In our comparison of these different solutions, it was found that the best regressor and cross-validation scores were obtained for the random forest regressor used in conjunction with the MutipleOutputRegressor.

Session 3: Cloud, HPC, and Hybrid Data Management

The third D3 session addressed bridging informational gaps between facilities, platforms, and databases. Presenters covered the approach to a joint DOE multi-facility data platform, how users can manage data for high-performance computing, and using cloud computing to upscale data management.

Challenges and Opportunities for a Joint DOE Nanoscale Science Research Centers Data Platform

Maria Chan | LANL

In FY18, the five DOE Nanoscale Science Research Centers (NSRCs) hosted 3,400 users, who make use of staff expertise and equipment in a variety of specialties, including nanomaterials synthesis; nanofabrication; electron, x-ray, and scanning probe microscopies; nanophotonics; nanobio materials; computational modeling; and more. According to Maria Chan of LANL in her plenary talk, unique to the NSRCs is the rich heterogeneity of user data emerging from the use of a wide range of instruments in synthesis, microscopy, spectroscopy, and computer simulations. At the present time, the different NSRCs have different schemes and different levels of implementation for acquiring, labeling, storing, and providing access to the heterogeneous data generated. To capture and curate data from the NSRCs, many levels of technical details need to be worked out, such as data formats, software development framework, sample tracking, metadata capturing, and labeling. In addition, datasets from correlated measurements, and the corresponding simulations, need to be handled in a coordinated manner to extract synergistic information. There is an urgent need for common standards and shared workflows for data across the NSRCs which will not only provide an effective data solution, but will also enable cross-center data sharing, augmentation, and manipulation. To achieve the promise of a FAIR (Findable, Accessible, Interoperable, Reusable) data ecosystem at the NSRCs, there is an urgent need to adopt a holistic approach that guides data, often in automated ways, from the point of acquisition through to the fully analyzed result. The presentation discussed how a coordinated data platform will change the status quo and allow users to significantly advance the use of their own data with annotation and analysis/AI/ML tools, and the reuse of shared, well-curated data for scientific discovery.

A User-Centered Data Management System for HPC

Annette Greiner and Lisa Gerhardt | LBNL

Wrangling data at a scientific computing center can be a major challenge for users, particularly when quotas may impact their ability to utilize resources. In such an environment, a task as simple as listing space usage for one's files can take hours, and sharing data efficiently requires specialized expertise, said Annette Greiner and Lisa Gerhardt of LBNL. To ease the pain of managing large data volumes, they designed and built a web-based data management system that allows users to easily manage their data while respecting the center's security policies. The data management system includes three interfaces for users. First, they designed and built a "Data Dashboard," a web-enabled visual application for the review of usage against quotas, discovering patterns, and identifying candidate files for archiving or deletion. A "PI Toolbox" is also in development to allow scientists to directly control the permissions of their files and directories and alleviate such common problems as updating millions of permissions. Systems are also gaining new users with less experience in transferring large files efficiently, leading to a "PB Data Portal" to facilitate sharing these large volumes of scientific data. The project is a new, generalizable framework on tools that come with common file system software, like Spectrum Scale, Lustre, and HPSS. Its user interfaces are built on common Javascript libraries (D3, React), with API calls to middleware scripts. It leverages file system scans, a Postgresql database, and Apache Spark for back-end metadata wrangling. The presentation described the process for developing tools, the framework supporting them, and the challenges for such a framework moving into the exascale age.

Evaluating Cloud Computing Capabilities to Support Scalable Data Management
Vic Baker | NETL

Scalable cluster computing is increasingly necessary due to the constantly increasing volume of historical and new data, according to Vic Baker of NETL. The ability to scale compute resources in a cost effective and timely manner is essential in compute intensive research. NETL has been evaluating Google Cloud Platform (GCP) through DOE HQ's instance using SmartSearch as the use-case. Evaluations of GCP focus on "lift and shift" GCP deployment analogous to NETL on-prem Hadoop deployments, as well as Kubernetes Engine, Cloud SQL, BigQuery, gcloud / gsutil terminal utils, and GCP ML capabilities. This talk covered lessons learned for both standalone GCP deployments as well as hybrid deployments between GCP and on-prem as they relate to devising and potentially deploying scalable, Cloud-enhanced solutions for data management and transformation needs of large and complex data stores.

Session 4: Data Access, Sharing, and Sensitivity

The fourth themed D3 session centered on access to data, with special consideration given to DOE security policy. Sessions tackled challenges relating to managing data through its long life, enabling many users at many different facilities to work with shared data, improving research collaborations, the collection and anonymization of data, streamlining data movement and access, and how to hold on to metadata after a dataset is fed to a machine learning model.

Fostering Data Curation Throughout the Entire Life Cycle of Energy Data Management
Chad Rowan | NETL

With the data revolution and FAIR data practices has come the recognition that scientific discovery through federally funded research products is limited to issues surrounding data curation and energy data management, said Chad Rowan of NETL in his plenary talk. In 2011, DOE's NETL began development and maintenance of the EDX to address the needs of data curation throughout the data life cycle while building the functionality needed to support a virtual laboratory. EDX supports the entire life cycle of data by securely sharing data from project inception to completion, facilitating and prudently governing secure access to team resources for multi-entity teams, and ultimately, ensuring preservation of that data and associated data products until the data is ready for publication. EDX utilizes a self-developed, highly customized version of CKAN to address the research needs associated with private sharing, in-house review of data products, and ultimately data publication with an accompanying data citation. EDX utilizes API connectivity, making published resources more easily discoverable. EDX supports NETL-affiliated research by coordinating historical and current data and information from a wide variety of sources to facilitate access to research that crosscuts multiple NETL projects/programs. EDX is underpinned by a robust governance protocol and procedure that addresses key challenges and needs. The platform hosts and, in some cases, virtualizes thousands of datasets encompassing millions of natural systems and engineering data features and attributes. The platform also has incorporated data visualization and virtual analytics through a web mapping application, Geocube, and containerized models and tools stemming from these research programs. In 2020, EDX released a major version 3 upgrade and was the recipient of the registered trademarks for EDX's name and logo by the United States Patent and Trademark Office (USPTO).

DOE Open Energy Data Initiative

Michael Rossol | NREL

Historically, the way public datasets from DOE and its national laboratories were accessed has been to download data from a website or data repository onto a personal computer, said Michael Rossol of NREL. While that approach works for small datasets (like those contained in Excel files) it is becoming untenable as the size of data continues to grow. The DOE Open Energy Data Initiative (OEDI) aims to improve and automate access of high-value energy data sets across the U.S. DOE's programs, offices, and national laboratories by partnering with AWS to build a cloud-based public datalake. The datalake will leverage AWS's public datasets program in conjunction with its state-of-the-art cloud computation resources to remove these barriers and improve accessibility for analysts and researchers. OEDI will enable data scientists and analysts to explore, mash-up, and analyze data in a framework that speeds innovation, allowing for rapid computation while also utilizing portions of their manipulated data for other purposes. The OEDI data catalog will document the contents of the datalake as well as other publicly available datasets that could be of used in conjunction with OEDIs public datasets.

Enhancing Research Team Collaborations in a Secure Environment with Energy Data eXchange (EDX) Drive

Daniel McFarland and Chad Rowan | NETL

Built atop the open-source data platform, CKAN, and tailored to adhere to DOE and federal data policies, orders and regulations, DOE's NETL EDX supports the full life cycle of data-driven research. Daniel McFarland and Chad Rowan of NETL explained the platform includes public data-resource curation capabilities for finalized products and collaborations. In 2018, EDX Drive was introduced to allow researchers to securely upload and manage files within EDX's private collaborative environments called Workspaces. Researchers can now privately share data resources such as PDFs, word documents, databases, software, and images, or provide external web URLs in a user interface similar to a desktop's file explorer. Files can be uploaded in bulk with drag-and-drop functionality and placed in subfolders for organizational purposes. Once uploaded, resource metadata can be edited to include descriptions and licenses to denote any associated restrictions and ensure end-users from multi-entity project teams are aware of key information to ensure appropriate use in a range of R&D efforts. Users within a Workspace may also review and provide ratings for a resource's completeness, source credibility, data consistency, and accessibility. Researchers in the Drive may bundle resources to create a submission to seamlessly initiate the public release review process from the Drive environment. Submissions are vetted and reviewed by NETL Project Managers or Team Supervisors, and once approved, they are released to the public with a DataCite.org formatted citation and other appropriate metadata for each matured product in compliance with FAIR standards. At present, EDX Drives across all the private workspaces account for around 75% of the data resources hosted by EDX; public, finalized assets account for 25%.

The Living Laboratory: Enabling Access to Operational Data for Research and Development

Elizabeth Jurrus | PNNL

A fundamental challenge towards the development of analytics and tools that aid in the analysis of complex energy, earth, and national security systems is access to representative datasets, said Elizabeth Jurrus of PNNL. As part of a scientific data sharing strategy, PNNL has created the Living Laboratory Data repository to provide access to real-world operational datasets for research and development. PNNL has created the Data

Stewardship Board (DSB) to oversee the collection of data, anonymization of that data, and agreements that manage the use of the data. Specifically, the DSB establishes mechanisms and guidance for the acceptance, use, and sharing of risk-sensitive data sets for PNNL research. The DSB addresses societal interests regarding business and privacy risks while also enabling the researchers to obtain, perform analysis, and share results in the pursuit of ethical science. The presentation served as an overview of the processes used for creating the Living Laboratory data, the Data Stewardship Board that manages and mitigates the risks associated with the data, and the research that the Living Laboratory enables. Most recently, the Living Laboratory has enabled a collaborative partnership between five DOE national laboratories focused on identifying patterns of interest from operational building data, combined with communications and travel information. Through the Living Laboratory, PNNL has also become the leading provider of data responsible for producing large-scale real-world background activity graphs that contain embedded activity pathways. Lastly, the data produced under the VOLTTRON program, included in the Living Laboratory, is used to enable researching analytics required to understand and predict complex relationships in the power grid.

Scalable Data Management for National Facilities Using the Modern Research Data Portal
Vas Vasiliadis | ANL

As data volumes grow, said Vas Vasiliadis of ANL, the research enterprise is increasingly challenged by what should be mundane tasks: reliably moving data from instruments and computing resources, easily describing data for downstream discovery, and making the data accessible (often with appropriate access controls) to distributed groups of collaborators. The presentation described common use cases in user facilities that motivate the need for such data portals, illustrated by further examples, and demonstrated how DOE investigators can rapidly develop and deploy these capabilities to scale up their research. One such example, among many, is the Petrel data portal (https://petreldata.net) developed by the ALCF and Globus, used by researchers to manage data in diverse fields including materials science, cosmology, machine learning, and serial crystallography. The portal facilitates automated ingest of data from APS beamlines and other sources, extraction and addition of metadata for creating search indexes, assignment of persistent identifiers faceted search for rapid data discovery, and point-and-click downloading of datasets by authorized users. The portal employs fine-grained permissions that control both visibility of metadata and access to the datasets themselves. It is based on the Modern Research Data Portal design pattern, jointly developed by the ESnet and Globus teams, and leverages capabilities such as the Science DMZ for enhanced performance and to streamline the user experience.

Kosh: An Open-Source Data Store for Large Datasets and Machine Learning Applications
Charles Doutriaux | LLNL

According to Charles Doutriaux of LLNL, ML problems are notoriously “hungry” for data. Even a “failed” simulation becomes most valuable as a teaching tool to the models. A difficulty for the ML data scientist is that, due to various reasons, such as the sensitive nature of their research, most projects work in an isolation bubble, with their own data storage formats and conventions. As a result, the data can be hard to discover and/or access by others, and codes can be very project specific. Kosh (“treasury” in Sanskrit) acts as an open-source centralized store in which data producers record their work with little to no overhead. The end-product is entered in a Kosh store as a “dataset.” Metadata (e.g., problem name, problem type, team, experiment name, code name, parameters used) is associated with the dataset to allow for easy (re)discovery later. Kosh’ Schema objects allow for validation of such metadata according to the project’s specific

conventions. In addition, data generated along this dataset is associated with it in Kosh. Kosh only stores a Uniform Resource Identifier (URI) pointing the data, its "MIME type" and associated metadata (e.g., data post-processed using some specific criteria). The MIME type is used by Kosh to match the data to a loader, allowing the end user to write similar code to retrieve very different data formats. Loaders (including custom data loaders) allow the data to be extracted to various output formats. Once a loader has been identified by Kosh, the data can possibly be further manipulated by Kosh's Transformers at extraction time. Kosh provides tools to easily move or copy data around while preserving accurate information in its stores.

Session 5: Data Curation Collection Specific Use Cases

Due to the idiosyncrasies and special concerns surrounding many DOE data regulation, additional talks were provided to address those special cases, including low-energy nuclear data, carbon storage data, and preserving legacy data that may be decades old.

Preparing for the Next 50 Years of Low-Energy Nuclear Data

Adam Hayes | BNL

Today, low-energy nuclear data (structure, decay, reactions, etc.) is typically stored in a set of text files, some in 1970s-era 80-column text format. Storing data in formatted files presents many difficulties, said Adam Hayes of BNL. Tremendous effort is put into updating formats, and updates to file formats can make them incompatible with codes that use them. Expanding the format to handle new types of data, sometimes even small additions, can be very difficult, and this discourages the inclusion of data that could be extremely useful to the nuclear physics community. While some research has been done into converting existing formats to hierarchical structure, such as XML, the contemporary paradigm is to use a true multipurpose database system to store data and to use files or serialization only to transmit data. This greatly simplifies and encourages the addition of more types of data to an existing database, in part because the addition of new data does not need to affect file formats in use. There are many types of new data that users could find extremely useful, such as binary experimental data sets, images, unpublished results, codes used in analysis, and open data. Arguments for a new approach based on object-oriented databases were made, based on a new database under development at the National Nuclear Data Center.

A Virtual Use Case for Carbon Storage Data Curation

Paige Morkner | NETL

Since 1997, the DOE has led an effort to implement geologic carbon sequestration demonstration projects and subsurface modeling tool development in the USA and parts of Canada, said Paige Morkner of NETL. As a result, large volumes of carbon storage data have been collected and produced, and in recent years, both the RCSPs and NRAP have transitioned to using the EDX (https://edx.netl.doe.gov) as the main platform for data storage and collaboration, with the goal of publicly publishing and hosting relevant datasets for researchers in the carbon storage community. DOE outlined the commitment to deliver federally funded data products to the public in 2014 with the publishing of the U.S. DOE Public Access Plan. In response to the publishing of large data volumes, there is now a need in the carbon storage community for intuitive data curation, labeling, and management strategies to enhance usability and discoverability. The presentation covered the methods and results of this curation effort and the tools developed to enhance future usability and discoverability of the data. The disparate nature of carbon storage data types applied to subsurface modeling, risk analysis and site screening presents a data curation challenge. Comprehensive data catalogs were developed for each

dataset to capture key metadata and data quality, and the spatial data density of 592 resources, consisting of over 630,000 attributes, was analyzed using the Cumulative Spatial Impact Layers tool (https://edx.netl.doe.gov/dataset/cumulative-spatial-impact-layers) to understand where data was and was not present. A set of virtual tools were established on EDX for enhanced data discoverability, including enhanced spatial search and keyword search capabilities able to integrate multiple data types (GeoCube; https://edx.netl.doe.gov/dataset/geocube), and an in-house-developed natural language processing (NLP) tool was created and applied for intuitive organization and keyword assignment of the text-based literature corpus of 2,071 documents.

DOE Office of Legacy Management Geospatial Data Lifecyle and Data Curation Project Development

Denise Bleakly | SNL

The DOE Office of Legacy Management (LM) is reviewing data lifecycle models for data curation concepts pertinent to their datasets, particularly geospatial data, said Denise Bleakly of SNL. Geospatial data are a special subset of digital data, which represent information tied to a location at the earth's surface or sub-surface. DOE LM has a charter to keep and manage data for at least 75 years into the future. Data curation is one of LM's main responsibilities so that data is discoverable and usable for generations to come. DOE LM is reviewing three data lifecycle models, the United States Geological Survey (USGS) "Science Data Lifecycle Model," the Federal Geographic Data Committee's "Stages of the Geospatial Data Lifecycle," and the Digital Curation Center "Data Curation Model." The USGS model emphasizes the process of collecting and managing scientific data, the Federal Geographic Data Committee (FGDC) model focuses on how data are managed based on business needs and the Data Curation Center's model focuses on data curation. Each of these data lifecycle models meet some of LM's data lifecycle needs, while none meet all. These three data lifecycle models will be reviewed and discussed in the context of geospatial data curation activities for DOE LM.

Hackathon Discussion

During the workshop, efforts were made to prepare for a virtual hackathon, which met with widespread support from workshop attendees. Participants were provided with a topic, instructions, and a dataset. Questions were assembled, and the 2020 workshop will serve as a seed for future Data Day-based hackathon work.

The hackathon challenge description was "Using an AI/ML based approach, automatically populate metadata fields in the simple schema that has been provided. Second, for extra credit, identify fields that would be useful to add to the simple schema."

Data is the new currency for the future and is critical to all DOE research programs. Data size and types are exponentially growing with advances in computational capability, experimental characterization capability, and a large, connected system of systems. The revolution of AI and ML are rapidly impacting all parts of science, but one thing remains the same – data scientists spend ~80% of their time curating and cleaning data rather than creating insights. Metadata is defined as *data about data* and makes it easier to retrieve, use, and generate insights from data. While numerous projects have developed project-specific metadata standards, we are not aware of an effort to develop a metadata schema that can be applied across a broad

range of data types or portfolios. Increasing standardization of metadata across DOE will provide a richer discovery capability and should accelerate the impact of AI and ML in discovering new scientific insights.

Participants were to choose at least one data set from the following at https://www.osti.gov/search/product-type:Data, selecting on *Research Org*:

- National Energy Technology Laboratory (NETL), Pittsburgh, PA, Morgantown, WV, and Albany, OR (United States). Energy Data eXchange (218)
- DOE Geothermal Data Repository (695)
- Environmental System Science Data Infrastructure for a Virtual Ecosystem (369)
- Oak Ridge National Lab. (ORNL), Oak Ridge, TN (US). Atmospheric Radiation Measurement (ARM) Data Center (296)
- Pacific Northwest National Lab. (PNNL), Richland, WA (United States). Atmosphere to Electrons (A2e) Data Archive and Portal

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.

Next-Gen Artificial Intelligence for Proliferation Detection

Nonproliferation remains one of the DOE's most important missions. For the DNN R&D Data Science Portfolio, developing the next generation of AI methods and technologies to detect early indicators of a foreign nuclear program's weapons-usable capabilities is critical. Commercially available AI is inadequate for the high-consequence missions of nuclear nonproliferation and therefore government R&D must close this gap. Through tight alignment with mission questions, government must drive R&D into new science and mathematics to overcome gaps where current capabilities fall short. Where appropriate, it is imperative to transition AI technologies to NNSA and mission partners that enhance government nuclear nonproliferation capabilities.

DOE Data Days Recap

DOE Data Days set the goal of promoting disciplined data management as a means to higher-quality and more efficient research and analysis, discussing data curation and standards; data-intensive computing and software tools; data access, sharing and sensitivity; and cloud, HPC, and hybrid data management. By bringing dispersed practitioners into the same space, Data Days can advance the standard practice and strengthen the community of practitioners across DOE.

During day 1, the overall conclusion was that metadata matters, a lot. The adoption of interoperable metadata schemas was a topic of discussion, in particular how to manage and promote these schemas. A common refrain was engaging in knowledge extraction: the mega-mining of data, metadata, text reports, or figures. Keeping communication lines open across the DOE complex may lead to the discovery of paths that will accelerate progress both in and out of the complex. Containers are proving useful for managing and processing large datasets, as is a wide variety of data management software from commercial and government vendors. Despite the growing importance of metadata and its management, however, stringent metadata protocols are proving counterproductive.

Presentations on day 2 focused on science applications. Data processing pipelines are enabling superior scientific progress alongside algorithms that can help pull useful information from noisy data streams. Work is underway on a number of multi-institutional research platforms that allow heterogeneous data to be shared seamlessly. Increasingly, infrastructure is focused on reducing the barrier to science with more automation and intuitive, user-centric dashboards for the entire workflow. Opportunities for next-generation data management include more automation, enhanced collaboration, and modularity-by-design.

On day 3, discussion continued on robust research platforms to share large-scale, heterogeneous data products. Attendees found value in exploring the overarching principles of FAIR. Given the sensitive nature of much of the DOE's work, presentations approached the sharing and use of anonymization and multi-permission data. Finally, a common theme continued to emerge about the complex's ability to scale its data management in the face of rapidly growing datasets and ever more detailed instruments.

DOE Data Days - Next Steps

The 2020 virtual edition of Data Days introduced a soft launch of Hackathon and a call for participants to reach out to the conference organizers. In late February, PNNL will lead a Next-Gen AI Workshop: Domain-Aware Methodologies.

As always, the organizing committee is always seeking feedback. One suggestion for future strategies was holding game theory-led virtual workshops. Accelerating next-gen AI for scientific impact requires next-gen data management and scientific computing advances, but people and policy barriers are proving much more challenging than technical barriers. While the goal remains proliferation detection, the potential for impact extends far beyond nuclear nonproliferation. The challenges posed by the nuclear security domain are so demanding that, in building AI to detect early nuclear proliferation, we will advance the entire field of AI.

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 FY21 data management hackathon event and grassroots organized brown bag seminars.

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 18% of whom replied. For more survey data, please see the Survey Results appendix.

Appendices

Organizing Committee

The second annual D3 event was organized by a multi-laboratory organizing committee representing many of the participating DOE laboratories.

Ghaleb Abdulla is with Lawrence Livermore National Laboratory.

Katherine (Kale) Anderson Aur is a seismologist and ground-based nuclear monitoring subject matter expert who specializes in large R&D, operational, and capability-based software projects at Sandia National Laboratories. As part of this role, she leads an effort to develop software for assessing the quality of real-time big data in addition to overseeing the execution of capabilities for the modernization of the U.S.'s ground-based nuclear explosion monitoring software. Furthermore, she also has metadata management expertise including participating in several data management teams for various R&D projects as well as developing standards and best practices for acquiring, assessing, archiving, and persisting data for real-time detection software, within large multi-disciplinary projects, and at the organizational level. Prior to working at Sandia, she worked at the IRIS PASSCAL seismic instrument center as a data specialist and a real-time systems analyst. Within that role, she served on an academic committee to improve the overall quality of all seismic data entering the IRIS DMC (the largest seismic data repository in the world), promoted tools to organize high-quality, research-ready, seismic datasets for the broad community and enhance quality-control feedback to seismic network operators.

Dr. Tammie Borders is the Technical Advisor for AI and Data Science in the Office of Proliferation Detection within the Defense Nuclear Nonproliferation Research & Development program at the National Nuclear Security Administration in the U.S. Department of Energy. Her home institution is Idaho National Laboratory, where she leads the Data and Software Sciences team, with a research portfolio utilizing artificial intelligence, advanced decision science frameworks, digital engineering, cloud architectures, and geospatial analytics. In 2019, she was selected as a national Diversity MBA Top 100 under 50 emerging leaders. Prior to joining INL, Dr. Borders worked in the defense industry on a number of research areas, including information fusion and threat characterization algorithms, computational materials informatics methodologies to accelerate nanotechnology-reinforced materials to the warfighter, and a variety of technology incubation projects, such as femtosecond laser technology for sensing and materials applications. She holds a Ph.D. in computational physical chemistry from the University of North Texas.

Jeffrey Burke is with the Kansas City National Security Campus.

Loni Cason is the Weapons and Complex Integration Principal Directorate Events Administrator with Lawrence Livermore National Laboratory. She has a B.S. in Business Administration from California State University, East Bay.

Shiloh Elliott is a modeling and simulation scientist in the National & Homeland Security Directorate at Idaho National Laboratory. She has a M.S. in Geographic Information Systems. Her research interest includes machine learning, decision support systems, graph-based dependency analysis, spatial analytics, and remote sensing.

Jessie Gaylord is the Division Leader for Global Security Computing Applications at Lawrence Livermore National Laboratory. In this role, she manages 170 computer scientists, data analysts, software engineers, and system architects supporting projects across program areas in Global Security, Climate, Bio, and Security and Protection. She also leads multi-phenomenology data collection on a series of large physics experiments at the Nevada National Security Site for the Low Yield Nuclear Monitoring Project, and previously held leadership roles on other multi-laboratory ventures supporting applied Data Science and data-intensive research for Defense Nuclear Nonproliferation. Ms. Gaylord initiated the annual DOE Data Days (D3) event and is very excited to see it continue to grow and evolve.

Dr. Daniel Laney is a computer scientist 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 WCI.

Angeline Lee is with Lawrence Livermore National Laboratory.

Gideon Juve is a software engineer at the Pacific Northwest National Laboratory specializing in data engineering and cloud computing for national security applications. Prior to joining PNNL, he developed systems for automating distributed computing applications at SpaceX and the University of Southern California Information Sciences Institute. He earned a Ph.D. in Computer Science from the University of Southern California.

Martin Klein is a scientist at Los Alamos National Laboratory.

Dr. M. Ross Kunz is a statistician for Idaho National Laboratory developing high-dimensional data visualization in 2D/3D environments and explainable AI techniques. His explainable AI work focuses on the fusion of machine learning and physics applied to a variety of tasks including chemical kinetics, nuclear process control, geology and electric vehicles. He has developed a 3-D visualization framework that allows emergency planners to simulate responses to various safety and security scenarios. His visualization has been presented at the White House and is now being used by federal, state and municipal leaders to plan for expanded use of electric vehicles. He holds a Ph.D. in statistics from Florida State University and a bachelor's in statistics from Idaho State University. Before joining INL in January 2015, he was a statistician for Michelin of North America.

Ruben Pino is with the Kansas City National Security Campus.

Christopher Ritter is a Group Lead with the Digital and Software Engineering group at Idaho National Laboratory. His expertise is in software engineering, software development, leading software teams, systems engineering software integration, and database management. Before coming to INL, he was director of software development at SPEC Innovations, in Manassas, Virginia. He served as the chief architect of Innoslate, a popular systems engineering tool that leverages elastic cloud technologies and AI/NLP for high scalability and advanced analytics. Architected the software system and consulted on the data ontology for a centralized mission risk management system for the Joint Staff at the Pentagon and supported Marine Corps

business process reengineering for its Capability Portfolio Management processes. He was also a computer programming teacher at St. Michael's Academy in Warrenton, Virginia, and developed an elementary school computer programming curriculum. He holds a bachelor's degree in computer science from Virginia Polytechnic Institute and State University.

Kelly Rose is a geology, geo-data science researcher with the National Energy Technology Laboratory's Research Innovation Center. Her research at NETL is focused on using geologic and geospatial science to reduce uncertainty about, characterize and understand spatial relationships between energy, engineered-natural systems at a range of scales. Her work involves development of new data-driven methods and tools for analysis of offshore energy, oil and gas, rare earth element, groundwater, carbon storage, and geothermal systems. Rose's research interests also include development of software driven solutions to common science-data curation, discovery and inter-operability challenges. She has served on advisory committees including the Department of Interior's National Geologic and Geophysical Data Preservation Program, United Nations Environmental Programme's global outlook on methane gas hydrates, and the University of Southern California's Induced Seismicity and Reservoir Monitoring Consortiums.

Dr. Stanley Ruppert, with a Ph.D. in seismology from Stanford University (1993), 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-laboratory ventures.

Angela Sheffield is the NA-22 Data Science Program Manager with the National Nuclear Security Administration.

Dr. Sandra Thompson is a group leader in the Global Nuclear Science and Technology group at Pacific Northwest National Laboratory. Her expertise is in data science for nuclear nonproliferation where she has built algorithms for remote sensing, instrumentation and information integration. She has led multi-laboratory, multi-disciplinary teams focused on solving mission problems, and continues to drive collaboration. She holds a Ph.D. in statistics from Colorado State University and a B.A. in mathematics from St. Olaf College.

Marc Wonders is an NNSA Graduate Fellow in NA-22. He received his Ph.D. in nuclear engineering from Pennsylvania State University and a B.S. from Washington and Lee University in Physics and Business Administration. His graduate research primarily focused on neutron detection and imaging for nonproliferation and security and includes published work on the application of silicon photomultipliers to new detectors, the development of novel scintillators, algorithm development, shielding simulations for electronic neutron generators, and signal readout techniques.

Attendees

Attendees represented a very diverse population of people from 27 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 1. Attendee organization as a percentage of total attendees (n=164).

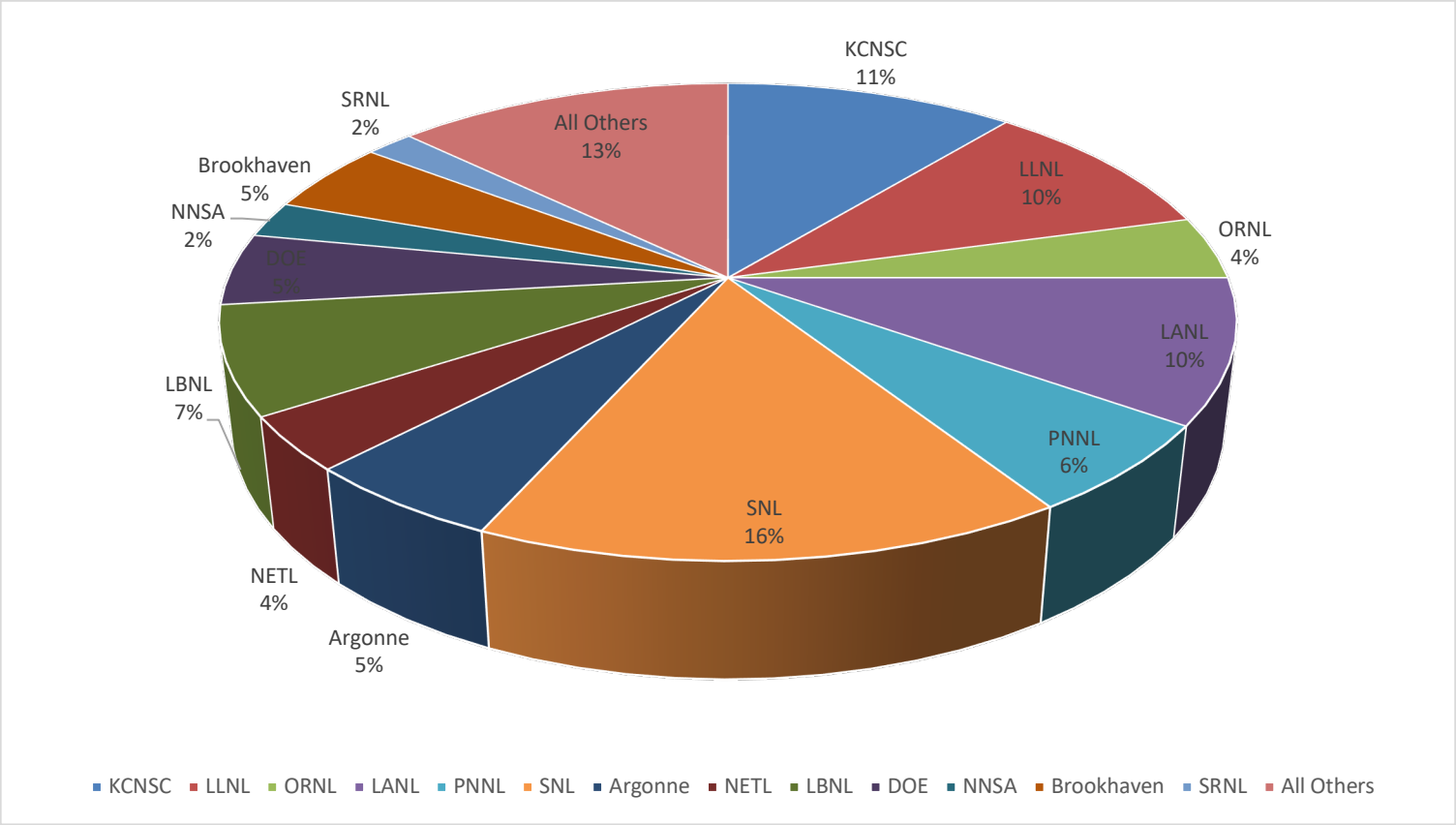

Figure 2. Concept and topic analysis derived from accepted abstract text. The text from each of the abstracts submitted to the workshop were analyzed using a trained, AI/ML natural language processing (NLP) algorithm to identify the frequency of words used within each abstract and ranked relative to the sessions they were aligned to. Stop words were removed from this analysis, focusing in on technical terms to help highlight themes of the workshop in general and of the topics. Below is a snippet of that analysis that highlights the top key words for each session topic.

3D Data Visualization		Data Access		Analytics		Data Model		Data Science Applications		Data Curation		Data Standards		ML Image Analysis		Big Data/HPC		Machine Learning	
Topic 0		**Topic 1**		**Topic 2**		**Topic 3**		**Topic 4**		**Topic 5**		**Topic 6**		**Topic 7**		**Topic 8**		**Topic 9**	
Word	**Propbability**	**Word**	**Propbability**	**Word**	**Propbability**	**Word**	**Propbability**	**Word**	**Propbability**	**Word**	**Propbability**	**Word**	**Propbability**	**Word**	**Propbability**	**Word**	**Propbability**	**Word**	**Propbability**
project	0.019	dataset	0.014	energy	0.025	use	0.013	use	0.015	data	0.018	metadata	0.018	use	0.019	use	0.017	model	0.028
metadata	0.018	data	0.013	standard	0.018	model	0.011	data	0.015	Data	0.015	standard	0.018	data	0.016	user	0.011	training	0.021
schema	0.016	platform	0.01	use	0.016	sensor	0.01	system	0.01	sample	0.013	fission	0.012	include	0.009	format	0.011	ML	0.017
nuclear	0.014	research	0.01	vehicle	0.014	network	0.01	user	0.009	use	0.013	data	0.011	EDX	0.008	network	0.011	use	0.011
NSRCs	0.011	resource	0.01	extension	0.014	high	0.01	develop	0.008	research	0.011	system	0.008	corridor	0.008	database	0.01	converter	0.011
use	0.01	Energy	0.009	estimate	0.012	system	0.009	model	0.008	scientific	0.01	GitHub	0.008	model	0.007	seismic	0.009	information	0.009
data	0.01	DOE	0.009	link	0.011	tool	0.009	storage	0.007	metadata	0.009	workflow	0.008	provide	0.007	time	0.009	performance	0.009
develop	0.009	hydropower	0.008	type	0.011	generate	0.009	analysis	0.006	Laboratory	0.008	community	0.008	resource	0.007	file	0.008	accuracy	0.009
research	0.009	support	0.008	consumption	0.011	dataset	0.009	source	0.006	user	0.008	repository	0.008	DOE	0.007	model	0.008	privacy	0.009

Figure 3. Attendee areas of expertise. Larger phrases in the word cloud indicate higher frequency.

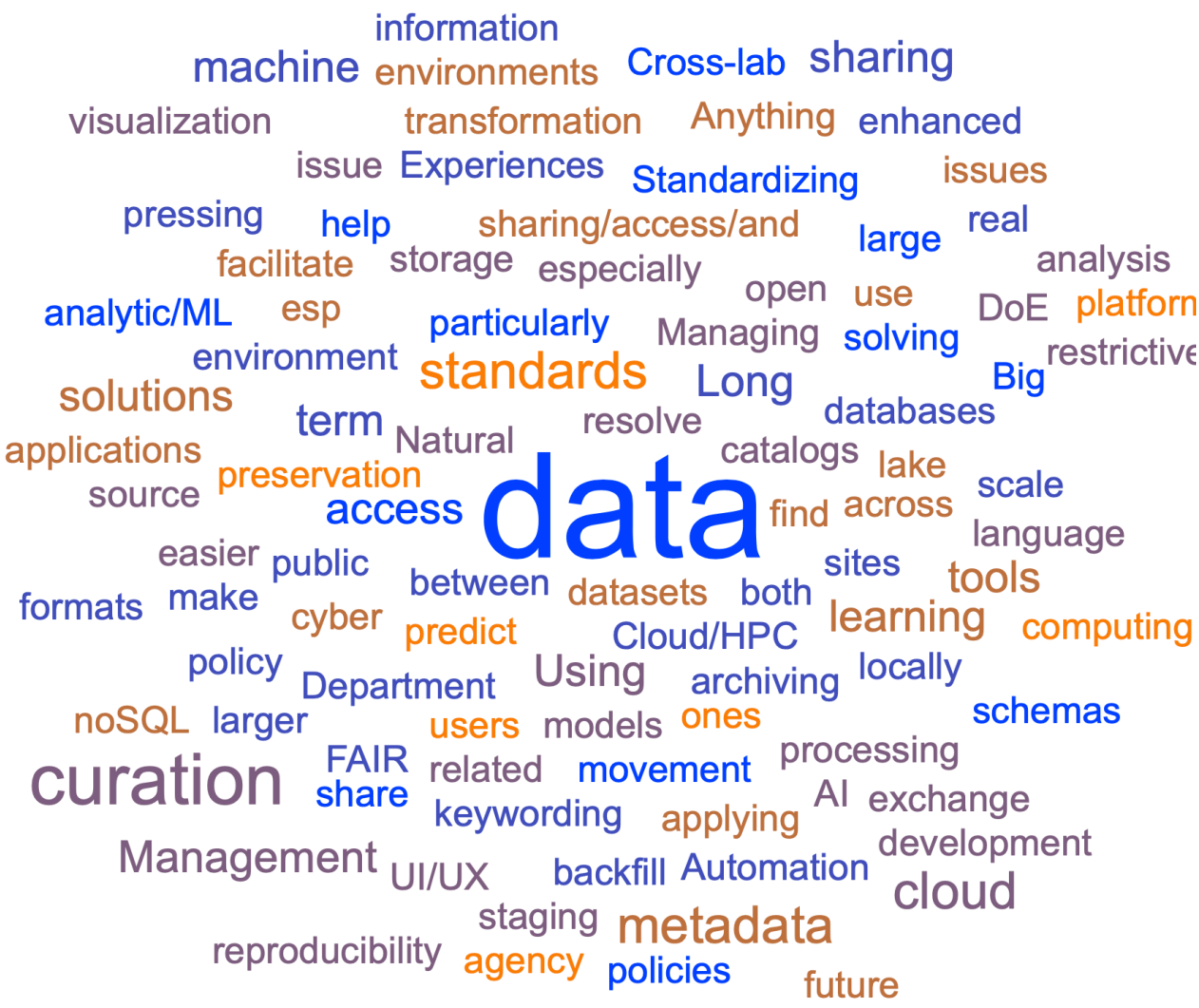

Figure 4. Attendee job titles. Larger phrases in the word cloud indicate higher frequency.

Attendees

1	Abraham, Ryan	rabraham@kcnsc.doe.gov	Kansas City National Security Campus
2	Acton, John	acton2@llnl.gov	Lawrence Livermore National Laboratory
3	Alam, Maksudul	alamm@ornl.gov	Oak Ridge National Laboratory
4	Alexandrov, Boian	boian@lanl.gov	Los Alamos National Laboratory
5	Almquist, Carrie	carrie.almquist@pnnl.gov	Pacific Northwest National Laboratory
6	Ames, Sasha	ames4@llnl.gov	Lawrence Livermore National Laboratory
7	Augustine, Joe	jaugustine@kcnsc.doe.gov	Kansas City National Security Campus
8	Aur, Katherine/Kale	kaaur@sandia.gov	Sandia National Laboratories
9	Avarca, Anthony	aavarca@anl.gov	Argonne National Laboratory
10	Baker, Vic	vic.baker@matricinnovates.com	NETL
11	Banks, Lawrence	banks12@llnl.gov	Lawrence Livermore National Laboratory
12	Bard, Deborah	djbard@lbl.gov	Lawrence Berkeley National Laboratory
13	Barnard, Edward	esbarnard@lbl.gov	Lawrence Berkeley National Laboratory
14	Bauer, Jennifer	jennifer.bauer@hq.doe.gov	DOE HQ
15	Berres, Anne	berresas@ornl.gov	Oak Ridge National Laboratory
16	Bleakly, Denise	drbleak@sandia.gov	Sandia National Laboratories
17	Borden, Rose	rmborde@sandia.gov	Sandia National Laboratories
18	Borders, Tammie	tammie.borders@nnsa.doe.gov	NNSA
19	Brown, David	dbrown@bnl.gov	Brookhaven National Laboratory
20	Brown, Elizabeth	ebrowndevkc@gmail.com	Masters Student at Georgia Tech
21	Burke, Jeffrey	jburke@kcnsc.doe.gov	Kansas City National Security Campus
22	Burke, Tami	Tami.burke@srnl.doe.gov	SRNL
23	Burrus, Madison	mburrus@lbl.gov	Lawrence Berkeley National Laboratory
24	Butterworth, Stefeni	butterworth6@llnl.gov	Lawrence Livermore National Laboratory
25	Byrnes, Susan	sbyrnes@sandia.gov	Sandia National Laboratories
26	Cain, Brian	cain_b@lanl.gov	Los Alamos National Laboratory
27	Calhoun, Don	dcalhoun@kcnsc.doe.gov	Kansas City National Security Campus
28	cappello, franck	cappello@anl.gov	Argonne national Laboratory
29	Chai, Chengping	chaic@ornl.gov	Oak Ridge National Laboratory
30	Chan, Maria	mchan@anl.gov	Argonne National Laboratory
31	Cook, Jeanine	jeacook@sandia.gov	Sandia National Laboratories
32	Cooke, Michael	michael.cooke@science.doe.gov	DOE OS
33	Crouch, Vickey	vcrouch@kcnsc.doe.gov	Kansas City National Security Campus
34	Crystal-Ornelas, Robert	rcrystalornelas@lbl.gov	Lawrence Berkeley National Laboratory
35	Cui, Helen	hhcui@lanl.gov	Los Alamos National Laboratory
36	Damerow, Joan	JoanDamerow@lbl.gov	Lawrence Berkeley National Laboratory

37	Damiani, Darin	darin.damiani@hq.doe.gov	DOE Office of Fossil Energy
38	Dehaan, Kristian	kdehaan@kcnsc.doe.gov	Kansas City National Security Campus
39	DeRaad, William	wderaad@sandia.gov	Sandia National Laboratories
40	Di, Sheng	sdi1@anl.gov	Argonne National Laboratory
41	Doutriaux, Charles	doutriaux1@llnl.gov	Lawrence Livermore National Laboratory
42	Downie, Carlos	downie4@llnl.gov	Lawrence Livermore National Laboratory
43	Elliott, Rory	roryelliott@lanl.gov	Los Alamos National Laboratory
44	Elsea, Stefanie	stefanie.elsea@cns.doe.gov	CNS Pantex
45	Ely, Kim	kely@bnl.gov	Brookhaven National Laboratory
46	Enders, Bjoern	benders@lbl.gov	Lawrence Berkeley National Laboratory
47	Everett, Maggie	meverett@kcnsc.doe.gov	Kansas City National Security Campus
48	Fagnan, Kjiersten	kmfagnan@lbl.gov	Lawrence Berkeley National Laboratory
49	Fisher, John	fisher@csail.mit.edu	MIT
50	Gaylord, Jessie	gaylord2@llnl.gov	Lawrence Livermore National Laboratory
51	Gerhardt, Lisa	lgerhardt@lbl.gov	Lawrence Berkeley National Laboratory
52	Gerics, Stephanie	gericss@osti.gov	OSTI
53	Goldsmith, Beth	goldsmith@lanl.gov	Los Alamos National Laboratory
54	Greiner, Annette	amgreiner@lbl.gov	Lawrence Berkeley National Laboratory
55	Hagengruber, Michael	mlhagen@sandia.gov	Sandia National Laboratories
56	Hanna, Craig	cjhanna@sandia.gov	Sandia National Laboratories
57	Harris, Ruth	raharri@sandia.gov	Sandia National Laboratories
58	Harvey, Dustin	harveydy@lanl.gov	Los Alamos National Laboratory
59	Havins, Shannon	shannon.havins@inl.gov	Idaho National Laboratory
60	Hayes, Adam	ahayes@bnl.gov	Brookhaven National Laboratory
61	Hennessy, Pat	phennessy@kcnsc.doe.gov	Kansas City National Security Campus
62	Hoang, Thuc	thuc.hoang@nnsa.doe.gov	NNSA
63	Hoellwarth, Loni	loni@llnl.gov	Lawrence Livermore National Laboratory
64	Huber, Cynthia	cmhuber@sandia.gov	Sandia National Laboratories
65	Huffer, Hillary	hillary.huffer@ferc.gov	FERC
66	Hughes, Hannah	hannah.hughes@ee.doe.gov	DOE EERE
67	Huitt, Drew	huittj@osti.gov	OSTI
68	Jackson, Stephen	sjacks@sandia.gov	Sandia National Laboratories
69	Jones, Tracy	tkjones@sandia.gov	Sandia National Laboratories
70	Joubert, Wayne	joubert@ornl.gov	Oak Ridge National Laboratory
71	Jurrus, Elizabeth	elizabeth.jurrus@pnnl.gov	Pacific Northwest National Laboratory
72	Juve, Gideon	gideon.juve@pnnl.gov	Pacific Northwest National Laboratory
73	Kiran, Mariam	mkiran@lbl.gov	Lawrence Berkeley National Laboratory
74	Klein, Martin	mklein@lanl.gov	Los Alamos National Laboratory
75	Kleinsorge, Kevin	kkleinsorge@kcnsc.doe.gov	Kansas City National Security Campus
76	Knapp, Doug	knapp22@llnl.gov	Lawrence Livermore National Laboratory

77	Land, Pam	pland@lanl.gov	Los Alamos National Laboratory
78	Laney, Daniel	laney1@llnl.gov	Lawrence Livermore National Laboratory
79	Lederman, Sol	ledermans@osti.gov	OSTI
80	Lewis, Jennifer	lewisje@sandia.gov	Sandia National Laboratories
81	Lofstead, Jay	gflofst@sandia.gov	Sandia National Laboratories
82	Lorek, Ryan	ryan.lorek@case.edu	Brookhaven National Laboratory
83	MacCarthy, Jonathan	jkmacc@lanl.gov	Los Alamos National Laboratory
84	Maceira, Monica	maceiram@ornl.gov	Oak Ridge National Laboratory
85	Madsen, Paul	pmadsen@kcnsc.doe.gov	Kansas City National Security Campus
86	Mahapatra, Sailendra	Sailendra.Mahapatra@hq.doe.gov	DOE HQ
87	Marcillo, Omar	marcillooe@ornl.gov	Oak Ridge National Laboratory
88	Maze, Julie	jmaze@lanl.gov	Los Alamos National Laboratory
89	McCutchan, Elizabeth	mccutchan@bnl.gov	Brookhaven National Laboratory
90	McKittrick, Alexis	alexis.mckittrick@ee.doe.gov	DOE EERE
91	Mendez, Jennifer	jen@pnnl.gov	Pacific Northwest National Laboratory
92	Miller, Laniece	lemiller@anl.gov	Argonne National Laboratory
93	Miramontes, Silvia	mirasilvia@lbl.gov	Lawrence Berkeley National Laboratory
94	Miranda, Raul	Raul.miranda@science.doe.gov	DOE OS
95	Mittrach, Michelle	mittrach@lanl.gov	Los Alamos National Laboratory
96	Morkner, Paige	Paige.Morkner@netl.doe.gov	NETL
97	Moyer, Elizabeth	emoyer@lanl.gov	Los Alamos National Laboratory
98	Nag, Ambarish	ambarish.nag@nrel.gov	NREL
99	Negron, Timothy	tnegron@kcnsc.doe.gov	Kansas City National Security Campus
100	Nelson, Joshua	nelsonjc@osti.gov	OSTI
101	Nicolae, Bogdan	bnicolae@anl.gov	Argonne National Laboratory
102	Nix, Kent	kent.nix@cns.doe.gov	CNS Pantex
103	Nobre, Gustavo	gnobre@bnl.gov	Brookhaven National Laboratory
104	Oblath, Noah	noah.oblath@pnnl.gov	Pacific Northwest National Laboratory
105	Obradovich, Joseph	joseph.obradovich@matricinnovates.com	NETL
106	Olaya Garcia, Paula Fernanda	polaya@vols.utk.edu	University of Tennessee, Knoxville
107	Orndorff, Gregory	gorndor@sandia.gov	Sandia National Laboratories
108	Otero, Pablo	paboter@sandia.gov	Sandia National Laboratories
109	Park, Gilchan	gpark@bnl.gov	Brookhaven National Laboratory
110	Pate, Russell	rdpate@sandia.gov	Sandia National Laboratories
111	Pearson, Matt	pearson31@llnl.gov	Lawrence Livermore National Laboratory
112	Perez, Jesus	jespere@sandia.gov	Sandia National Laboratories

113	Peterson, Ryus	rpeterson@kcnsc.doe.gov	Kansas City National Security Campus
114	Pew, Dallin	pewdc@nv.doe.gov	MSTS
115	Pike, Jeff	jeff.pike@srnl.doe.gov	SRNL
116	Plapp, Brendan	brendan.plapp@hq.dhs.gov	DHS
117	Pope, Paul	papope@lanl.gov	Los Alamos National Laboratory
118	Pouchard, Line	pouchard@bnl.gov	Brookhaven National Laboratory
119	Prout, Ryan	proutrc@ornl.gov	Oak Ridge National Laboratory
120	Purohit, Sumit	Sumit.Purohit@pnnl.gov	Pacific Northwest National Laboratory
121	Quiter, Brian	bjquiter@lbl.gov	Lawrence Berkeley National Laboratory
122	Ramprakash, Sreeranjani	jini@anl.gov	Argonne National Laboratory
123	Richardson, Steve	steve.richardson@nrl.navy.mil	US Naval Research Laboratory
124	Roberts, Herbert	hroberts@kcnsc.doe.gov	Kansas City National Security Campus
125	Rodd, Rebecca	rodd2@llnl.gov	Lawrence Livermore National Laboratory
126	Rose, Kelly	kelly.rose@netl.doe.gov	NETL
127	Rossol, Michael	michael.rossol@nrel.gov	NREL
128	Rowan, Chad	chad.rowan@yahoo.com	NETL
129	Ruppert, Stanley	ruppert1@llnl.gov	Lawrence Livermore National Laboratory
130	Russell, Thomas	thomas.russell@science.doe.gov	DOE OS
131	Salmond, Josh	jsalmond@kcnsc.doe.gov	Kansas City National Security Campus
132	Schoch, Dave	dgschoc@sandia.gov	Sandia National Laboratories
133	Schwarz, Nicholas	nschwarz@anl.gov	Argonne National Laboratory
134	Senter, Lee	senterlm@nv.doe.gov	MSTS
135	Sheffield, Angela	angela.sheffield@nnsa.doe.gov	NNSA
136	Sims, Benjamin	bsims@lanl.gov	Los Alamos National Laboratory
137	Sjaardema, Gregory	gdsjaar@sandia.gov	Sandia National Laboratories
138	Smith, Emily	esmith1@ameslab.gov	Ames Laboratory
139	Smith, Janice	jjsmit@sandia.gov	Sandia National Laboratories
140	Stearman, Terri	stearman3@llnl.gov	Lawrence Livermore National Laboratory
141	Stephan, Eric	eric.stephan@pnnl.gov	Pacific Northwest National Laboratory
142	Suckow, Thomas	Thomas.Suckow@pnnl.gov	Pacific Northwest National Laboratory
143	Tallent, Nathan	tallent@pnnl.gov	Pacific Northwest National Laboratory
144	Talley, Kevin	KevinRTalley@Gmail.com	NREL
145	Taylor, Nicholas	ntay@lanl.gov	Los Alamos National Laboratory
146	Teranishi, Keita	knteran@sandia.gov	Sandia National Laboratories
147	Thompson, Erich	ethompson@kcnsc.doe.gov	Kansas City National Security Campus
148	Toomey, John	jtoomey@sandia.gov	Sandia National Laboratories

149	Trujillo, Joshua	jtrujillo@kcnsc.doe.gov	Kansas City National Security Campus
150	Ulmer, Craig	cdulmer@sandia.gov	Sandia National Laboratories
151	Vasiliadis, Vas	vasv@anl.gov	Argonne National Laboratory
152	Venezuela, Otto	venezuela1@llnl.gov	Lawrence Livermore National Laboratory
153	Vo, Tom	vo13@llnl.gov	Lawrence Livermore National Laboratory
154	Watson, William	watsonw@osti.gov	OSTI
155	Wendell, Kathleen	kwendell@kcnsc.doe.gov	Kansas City National Security Campus
156	Wenzlick, Madison	madison.wenzlick@netl.doe.gov	NETL
157	West, Mary Beth	westm@osti.gov	OSTI
158	Wheeler, Lauren	lwheele@sandia.gov	Sandia National Laboratories
159	Whitlock, Daren	dwhitlock@kcnsc.doe.gov	Kansas City National Security Campus
160	Wonders, Marc	marc.wonders@nnsa.doe.gov	NNSA
161	Wong, James	jim.wong@srs.gov	SRNL
162	Wood, Lynn	lynn.wood@pnnl.gov	Pacific Northwest National Laboratory
163	Yeager, Chris	cyeager@lanl.gov	Los Alamos National Laboratory
164	Young, Brian	byoung@sandia.gov	Sandia National Laboratories

Survey Results

Of the 164 event participants, 19 completed the online survey for a 11% response rate, down from the previous year. 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, while face-to-face interactions are preferred, welcomed the virtual format during COVID-19. The overall sentiment shared by most respondents was that D3 was a valuable event with a necessary future.

Sixty-eight percent of the 19 respondents were "very satisfied" with D3 overall, while 26% were "satisfied" and one participant was "neutral." All participants were "satisfied" or "very satisfied" with the presentations, and no respondents were "neutral" or "dissatisfied." Only three attendees felt "neutral" about the 2020 virtual format, with the overwhelming majority (84%) "satisfied" or "very satisfied."

The range of topics at the 2020 D3 were mostly "satisfying" or "very satisfying" to attendees (89%), with one attendee "neutral" and another "dissatisfied." However, 100% of attendees were "satisfied" or "very satisfied" with the quality of all topics.

Attendees rated the following topics as most beneficial: machine learning and AI; data curation and standards; curated DOE repositories; data access, sharing and sensitivity; catalogs and interoperability; data management; metadata schemas; FAIR principles; and cloud services. Some participants noted interest in a broad range of the speakers' techniques and systems, as well as on the DOE JGI and OSTI.

Over 50% of respondents were "satisfied" or "very satisfied" with the hackathon concept, while one respondent was "dissatisfied," two were "neutral," and five did not respond. The moderated Q&A forums at the end of each session were largely satisfying to participants, with 84% responding favorably.

When asked what they liked most about D3, respondents cited the diversity of presentations, the broad spectrum and combination of data management and analysis topic areas, the flexibility of tuning into specific presentations of interest, the opportunity to make new contacts for collaboration, and the ability to watch recorded presentations after sessions were over. One respondent appreciated not having to travel to attend.

As per the survey respondents, future D3 sessions — unanimously encouraged — should include a forum for further collaborations, increased organization between sessions, additional interactivity, and a live chat function (instead of a time-delayed Q&A). One respondent requested that the 2021 event not include COVID. In the future, attendees would like to see sessions on streaming data management and analysis, applying analytic tools to large datasets, navigating policy roadblocks, natural language processing, machine learning, AI and industry tools, portable data science workflows, DOE cloud policies, data scales, and database organization strategies. The October timeframe for D3 was criticized for being too close to the end of the fiscal year, and many respondents suggested holding the next event in the spring or summer. One suggestion

was to have smaller "brown-bag" events throughout the year on D3 topics. Two-thirds of respondents planned to collaborate with other DOE sites in the future after D3, explaining that now they were aware of more people in relevant fields to work with.

Respondents found out about D3 in a number of ways: from their coworkers and managers, from emails, word-of-mouth, past presenters, members of the steering committee, other organizations around the DOE and the NA-22 website.

Other selected participant comments:

"Very well run workshop! Kudos to the organizing committee. I wish there was a good way to also interact with other attendees of this event, though. Perhaps that should be part of future workshops?"

"The fraction of women presenting and moderating was surprisingly high. I hope to see this again."

Acronyms

Acronym	Definition
AI	Artificial intelligence
ALCF	Argonne Leadership Computing Facility
ANL	Argonne National Laboratory
API	application programming interface
APS	Advanced Photon Source
ASCR	Advanced Scientific Computing Research
AWS	Amazon Web Services
BNL	Brookhaven National Laboratory
D3	DOE Data Days
DAQ	data acquisition
DNN	Defense Nuclear Nonproliferation
DOE	Department of Energy
DS	Data Science
DSB	Data Stewardship Board
EDX	Energy Data eXchange
EMN	Energy Materials Network
ENSDF	Evaluated Nuclear Structure Data File

Acronym	Definition
ESS-DIVE	Environmental Systems Science Data Infrastructure for a Virtual Ecosystem
FAIR	Findable, Accessible, Interoperable, Reusable
FGDC	Federal Geographic Data Committee
FY	fiscal year
GCP	Google Cloud Platform
GMP	Geophysical Monitoring Program
HEP	high energy physics
HPC	high performance computing
HTEM	High Throughput Experimental Materials
HQ	Headquarters
INL	Idaho National Laboratory
IT	information technology
JGI	Joint Genome Institute
KCNSC	Kansas City Nuclear Security Campus
kKNN	k-Nearest Neighbors
LANL	Los Alamos National Laboratory
LBNL	Lawrence Berkeley National Laboratory
LLNL	Lawrence Livermore National Laboratory
LM	Office of Legacy Management
ML	machine learning
NA-22	Nonproliferation Research and Development
NASA	National Aeronautics and Space Administration
NERSC	National Energy Research Scientific Computing Center
NETL	National Energy Technology Laboratory
NIST	National Institute of Standards and Technology
NLP	natural language processing

Acronym	Definition
NNSA	National Nuclear Security Administration
NREL	National Renewable Energy Laboratory
NSF	National Science Foundation
NSLS-II	National Synchrotron Light Source II
NSRC	Nanoscale Science Research Center
OEDI	Open Energy Data Initiative
OSTI	Office of Scientific and Technical Information
PNNL	Pacific Northwest National Laboratory
Q&A	question and answer
R&D	research and development
REST	representational state transfer
RF	random forest
S&T	science and technology
SNL	Sandia National Laboratories
SSX	serial synchrotron crystallography
TAZeR	Transparent Asynchronous Zero-copy Remote I/O
URI	Uniform Resource Identifier
USGS	United States Geological Survey
USPTO	United States Patent and Trademark Office
UT Knoxville	University of Tennessee, Knoxville
WCI	Weapons and Complex Integration Principal Directorate
XAS	X-ray Absorption Spectroscopy
XPCS	X-ray Photon Correlation Spectroscopy