DATA AND MODELS:

A FRAMEWORK FOR ADVANCING AI IN SCIENCE

Report of the Office of Science Roundtable on Data for AI | June 5, 2019

U.S. DEPARTMENT OF ENERGY | Office of Science

Data and Models: A Framework for Advancing AI in Science

Report of the Office of Science Roundtable on Data for AI

Report co-authors:

Name	Institution
Kjiersten Fagnan	Lawrence Berkeley National Laboratory
Youssef Nashed	Argonne National Laboratory
Gabriel Perdue	Fermi National Accelerator Laboratory
Daniel Ratner	SLAC National Laboratory
Arjun Shankar	Oak Ridge National Laboratory
Shinjae Yoo	Brookhaven National Laboratory

Sponsored by:
U.S. Department of Energy, Office of Science
DOE Office of Science Technical Contact: Laura Biven (*Laura.Biven@science.doe.gov*)

Fagnan, Kjiersten, Nashed, Youssef, Perdue, Gabriel, Ratner, Daniel, Shankar, Arjun, and Yoo, Shinjae. *Data and Models: A Framework for Advancing AI in Science*. United States: N. p., 2019. Web. doi: 10.2172/1579323

Contents

Figures

Executive Summary

On June 5, 2019, the Office of Science (SC) organized a one-day roundtable to focus on enhancing access to high-quality and fully traceable research data, models, and computing resources to increase the value of such resources for artificial intelligence (AI) research and development and the SC mission.[1] In this report, we consider AI to be inclusive of, for example, machine learning (ML), deep learning (DL), neural networks (NN), computer vision, and natural language processing (NLP). We consider "data for AI" to mean the digital artifacts used to generate AI models and/or employed in combination with AI models during inference. In part, this roundtable was motivated by the recognition that a large portion of science data currently are not well suited for AI.

The roundtable participants represented expertise from 12 Department of Energy (DOE) national labs[2], the National Institutes of Health (NIH), and the National Science Foundation (NSF), and they had wide-ranging knowledge in areas spanning the domain sciences, as well as from AI, data management, data curation, metadata, library sciences, storage systems and input/output (I/O), open data, big data, and edge computing. These experts also represented mission drivers across the six SC programs[3] and the DOE Office of Scientific and Technical Information (OSTI) with ties to SC-supported research activities, scientific user facilities, and community data repositories.

The fundamental finding of this roundtable is that there are opportunities to advance AI research and development (R&D) and increase the benefit of AI to science by improving the reusability of science data and AI models and through the development of methodologies and services to seamlessly and routinely integrate AI into science workflows. The roundtable participants identified three priority opportunities for data to advance AI in science:

1) Influence the development of AI tools by democratizing access to benchmark science data
2) Make AI operational in science with composable services for simulation, data analysis, and AI at all scales
3) Address open questions in AI with frameworks for relating data, models, and tasks.

These **opportunities** are presented in a broader context of open research **challenges** in AI and prerequisite, enabling **capabilities** in data science and data management.

---

[1] The focus for the roundtable is motivated by the Executive Order on Maintaining American Leadership in Artificial Intelligence, Feb. 11, 2019. https://www.whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/

[2] Argonne National Laboratory, Brookhaven National Laboratory, Fermi National Accelerator Laboratory, Lawrence Berkeley National Laboratory, Lawrence Livermore National Laboratory, Los Alamos National Laboratory, Oak Ridge National Laboratory, Pacific Northwest National Laboratory, Princeton Plasma Physics Laboratory, Sandia National Laboratory, SLAC National Accelerator Laboratory, Thomas Jefferson National Accelerator Facility.

[3] The six Office of Science programs are: Advanced Scientific Computing Research (ASCR); Basic Energy Sciences (BES); Biological and Environmental Research (BER), Fusion Energy Sciences (FES), High Energy Physics (HEP), and Nuclear Physics (NP).

1. Introduction

The U.S. Department of Energy (DOE) Office of Science (SC) has a unique combination of capabilities to lead the nation in artificial intelligence (AI) and machine learning (ML) research and development (R&D) for science:

- A broad mission that presents new and unique research problems on a national and global scale to attract new talent
- Sources of massive and/or complex science and engineering data from sensors, instruments, and from large-scale simulations
- World-class high-performance computing (HPC) infrastructure, capable of world-leading AI research
- World-class high-performance network infrastructure capable of integrating computing resources and data assets
- An exceptional workforce with large numbers of domain scientists, computer scientists, and mathematicians currently engaged in AI and related fields.

Current efforts, however, are impeded by difficulties in finding, accessing, preparing, sharing, reusing, and computing on science data. Researchers who want to develop new AI algorithms and techniques rely on data that are readily available and, preferably, curated with relevant metadata. Once generated, potential training datasets and models may go underutilized due to the lack of sharing platforms and practices, difficulty of moving or accessing data, and complexity in preparing data for computation. These challenges are particularly acute for the DOE SC because of the extreme scale and complexity of the data and, for many disciplines, the lack of established repositories and tools to facilitate best practices in data management, sharing, and preservation. Most of the data produced through SC-funded research are not used in AI applications, resulting in missed opportunities to use AI to make science more efficient and productive or to attract talent to the DOE mission in highly competitive AI areas. The SC mission and workforce are likely to benefit from a strategic approach to data for AI.

On June 5, 2019, the SC organized a one-day roundtable to focus on enhancing access to high-quality and fully traceable research data, models, and computing resources to increase the value of such resources for AI R&D and the SC mission.[4] In this report, we consider AI to be inclusive of, for example, ML, deep learning (DL), neural networks (NN), computer vision, and natural language processing (NLP). We consider "data for AI" to mean the digital artifacts used to generate AI models and/or used in combination with AI models during inference. In part, this roundtable was motivated by the recognition that a large portion of science data currently are not well suited for AI. The roundtable participants represent expertise from 12 DOE national labs[5], the National Institutes of Health (NIH), and the National Science Foundation (NSF).

[4] The focus for the roundtable is motivated by the Executive Order on Maintaining American Leadership in Artificial Intelligence, Feb. 11, 2019. https://www.whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/

[5] Argonne National Laboratory, Brookhaven National Laboratory, Fermi National Accelerator Laboratory, Lawrence Berkeley National Laboratory, Lawrence Livermore National Laboratory, Los Alamos National Laboratory, Oak Ridge National Laboratory, Pacific Northwest National Laboratory, Princeton Plasma

Participants' expertise spanned the domain sciences to AI, data management, data curation, metadata, library sciences, storage systems and input/output (I/O), open data, big data, and edge computing. These experts represented mission drivers across the six SC programs and the Office of Scientific and Technical Information (OSTI) with ties to SC-supported research activities, scientific user facilities, and community data repositories.

In January 2018, the ASCR Basic Research Needs workshop on Scientific Machine Learning,[6] was convened to identify major ML opportunities and grand challenges as viewed through the lens of applied mathematics and scientific computing research. That workshop identified six Priority Research Directions (PRDs) for Scientific Machine Learning. The first three PRDs describe foundational research themes common to the development of all Scientific Machine Learning methods and correspond to the need for domain-awareness (PRD #1), interpretability (PRD #2), and robustness (PRD #3). The other three PRDs describe capability research themes and correspond to the three major use cases of Scientific Machine Learning for massive scientific data analysis (PRD #4), ML-enhanced modeling and simulation (PRD #5), and intelligent automation and decision-support of complex systems (PRD #6). Together, these PRDs define the SC research goals for ML (PRDs 1-3) and provide broad classes of use cases where ML can impact the SC mission (PRDs 4-6). As a complement to the Scientific Machine Learning workshop, this roundtable focused on the opportunities and challenges related to data needed to advance these areas of research, as well as to advance the impact of AI and ML on the SC mission. The roundtable participants carefully considered the role of data in modern science applications of AI and in AI R&D, including issues around data generation and curation.

The roundtable began with discussions around how to make science data FAIR7 (Findable, Accessible, Interoperable, and Reusable) for AI (Figure 1). The creation of FAIR, annotated training data currently requires human expertise and curation. There are ongoing discussions in various science communities about the actual processes and feasibility of enabling data sharing and implementing the FAIR[8] principles.

Metadata and standards are key enablers of FAIR data, and, throughout the roundtable, these topics were at the heart of many discussions. Standards are recognized as powerful enablers of FAIR data. However, they must be designed carefully to avoid limiting systems to narrow syntax or semantics. Furthermore, most of the currently defined standards are not used consistently, and it is challenging to enforce their use outside of large repositories.

Physics Laboratory, Sandia National Laboratory, SLAC National Accelerator Laboratory, Thomas Jefferson National Accelerator Facility.

[6] Baker, Nathan, Alexander, Frank, Bremer, Timo, Hagberg, Aric, Kevrekidis, Yannis, Najm, Habib, Parashar, Manish, Patra, Abani, Sethian, James, Wild, Stefan, Willcox, Karen, and Lee, Steven. Workshop Report on Basic Research Needs for Scientific Machine Learning: Core Technologies for Artificial Intelligence. United States: N. p., 2019. Web. doi:10.2172/1478744.

[7] Wilkinson, M. D. et al. The FAIR Guiding Principles for Scientific Data Management and Stewardship. Sci. Data 3:160018 doi: 10.1038/sdata.2016.18 (2016).

[8] e.g., Research Data Alliance, https://www.rd-alliance.org/, and GO-FAIR https://www.go-fair.org/ initiatives.

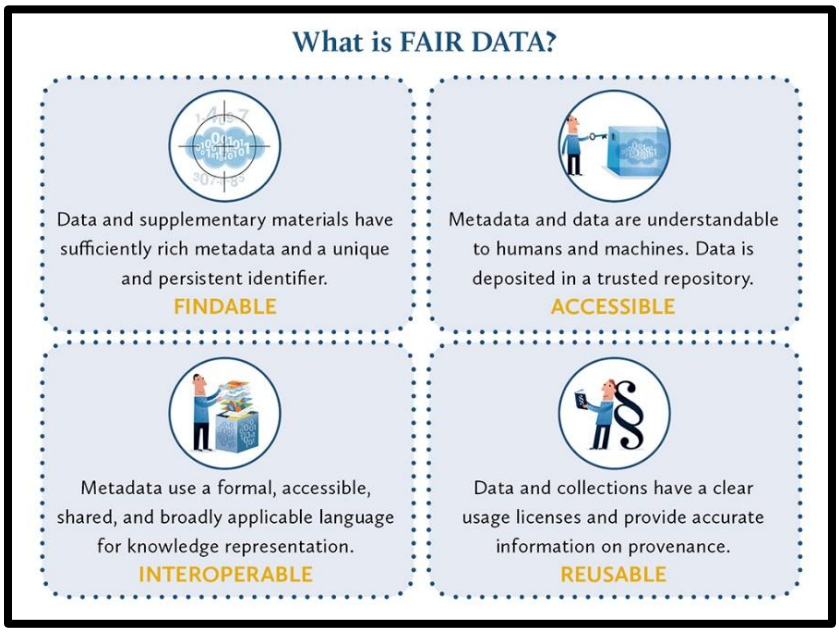

Figure 1: Graphical depiction of FAIR data principles from LIBER, the Association of European Research Libraries (https://libereurope.eu/wp-content/uploads/2017/12/LIBER-FAIR-Data.pdf).

The findability, accessibility, interoperability, and reusability of AI models emerged in the roundtable discussions as important considerations. An AI model is an inference method that can be used to perform a "task," such as prediction, diagnosis, classification, etc. Conceptually, AI models fall somewhere between data and theory—neither entirely empirical nor wholly derived from first principles. When used for control or autonomous decision making in a scientific workflow, the trained model may be an important digital artifact for reproducibility of the results. It also may be an important element of provenance for the resulting scientific dataset. In other cases, the model may be better viewed as an approximation of data—either the data used to train it or the data it generates, as in the case of Generative Adversarial Networks (GANs). However, as digital research objects, AI models are in their infancy with very few schemas for syntax, ontologies (or even controlled vocabulary), or metadata standards.

The fundamental finding of this roundtable is that there are opportunities to advance AI R&D and increase the benefit of AI to science by improving the reusability of science data and AI models and through the development of methodologies and services to seamlessly and routinely integrate AI into science workflows. These opportunities are presented in a broader context of open research challenges in AI and prerequisite, enabling capabilities in data science and data management (Figure 2). This report follows the Figure 2 structure by addressing each of the framework elements in turn in the next section entitled, *Findings*.
Section 2.1 provides details on the four open **challenges** in AI highlighted by the roundtable participants:

- Current AI tools and methods are not always a good fit for science.
- Science workflows with AI need human oversight.

- There is no theory encompassing data, AI models, and tasks.
- Applying FAIR principles to science data is challenging.

Section 2.2 presents each of the key priority **opportunities** for data to advance AI for science:

- Influence the development of AI tools by democratizing access to benchmark science data
- Make AI operational in science with composable[9] services for simulation, data analysis, and AI at all scales
- Address open questions in AI with frameworks for relating data, models, and tasks.

Section 2.3 describes underlying **capabilities** in data science and data management needed to address the key priority opportunities:

- Data management support and incentives for teams generating data
- Automated collection of metadata, provenance, and annotations at scale
- Scalable, human interfaces for data
- Strategic approaches to managing data management costs and resources.

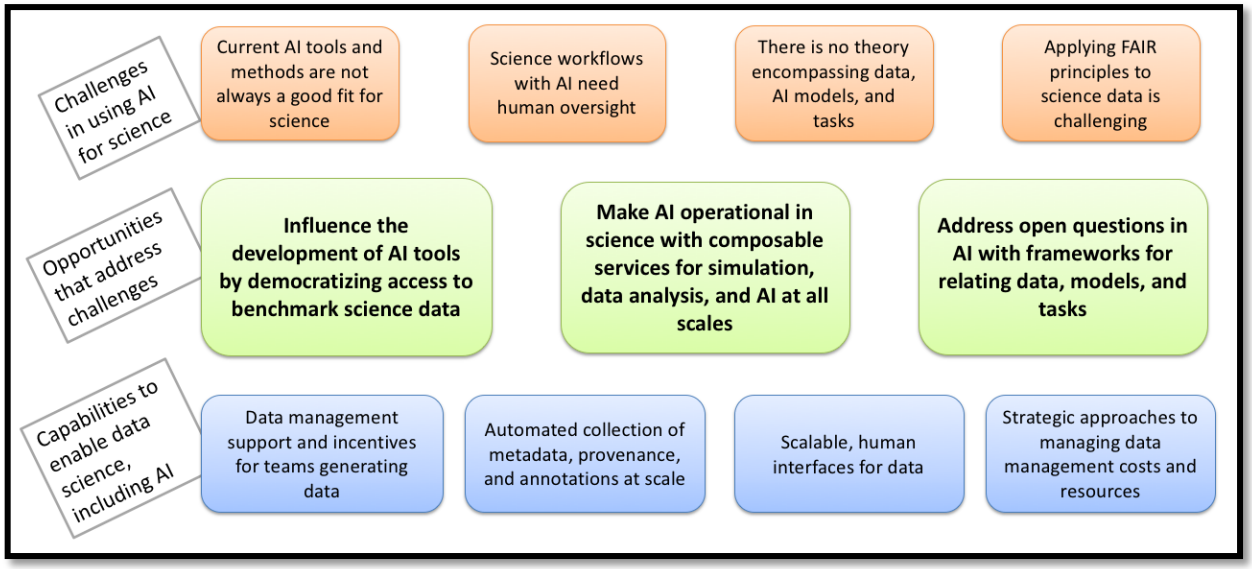

Figure 2: Open challenges in AI (top); opportunities to address these through data (middle); and prerequisite, enabling capabilities in data science and data management (bottom).

2. Findings

2.1 Challenges

The roundtable participants identified several current challenges to advancing AI R&D and using AI to advance the DOE SC mission. These challenges point to needs in technology

[9] Composable services are created from interoperable modular components that can be assembled flexibly into multiple well-defined functional and usable tools or capabilities.

development; fundamental research; and new types of collaborations that bring together experts in domain science, AI, and data management.

2.1.1 Current AI tools and methods are not always a good fit for science

Roundtable participants identified three reasons why current AI tools are not always a good fit for science use cases: 1) because the data used to develop them differ from science data in fundamental ways, 2) because science applications of AI can have goals that differ from traditional AI tasks, and 3) because of the often extreme conditions under which AI is used in science.

Advances in AI have exploited readily available data from images, sound, natural language, and game-like environments. However, science data can be high-dimensional, multimodal, complex, structured, and/or sparse. AI tools are sensitive to data representations, for example, differences in how data are sampled, averaged, or organized. Therefore, it is not clear about how to represent and organize science data for AI applications. In general, the distinguishing features, dependencies, and fundamental relationships within and among science data mean that traditional AI tools, which are typically designed for discrete, combinatorial, and unstructured data representations and analysis, may miss the scientific phenomena of interest.

SC-funded research generates large amounts of science data across a variety of domains through HPC simulations and experiments, many of which are single-facility multimodal experiments. Examples include the “onion”-style detectors of high energy physics and nuclear physics accelerator-based experiments, which have long integrated diverse detector types to measure different particle types and parameters. Increasingly, photon science experiments also combine multiple sources, for example, combining photon, ion, and electron yields in chemical studies. At X-ray free electron laser facilities (XFELs), users commonly integrate sample measurements with facility diagnostics to clean, correct, or enhance datasets. The Nanoscale Science Research Centers (NSRCs) also employ multimodal experiments. In these cases, data are generated through multimodal means from a variety of sensors and often are correlated in space and time, reflecting the underlying structures and processes being examined. The data usually are high-dimensional as well because of the large number of parameters needed to specify the experimental conditions and system state. Once synthesized into physical events or traces, data can be sparse with respect to learning task categories. Another layer of multimodality can result when data from multiple experiments or simulations are combined in a single AI application.

Error, uncertainty, and resolution add complexity to multimodal science data. Heterogeneous data sources have variable error sources, uncertainty levels, and resolution. For instance, consider the seemingly straightforward task of combining simulation and experimental data, which have different types and sources of uncertainty and error. At a minimum, confidence levels and error sources need to be captured explicitly as parameters or contained in metadata.

Science applications of AI can have different goals than traditional AI tasks and may pursue different data management challenges. Current AI capabilities that appear to dramatically

outperform human intelligence include recognizing recurring patterns in images, sounds, natural language, and game-like environments. Scientists, however, are interested in tasks not necessarily modeled on human intelligence such as identifying atypical and anomalous cases—often under extreme computational environments of large, complex data; extreme data rates; and low latency tolerances. There are significant associated challenges in data management, data movement, and data preparation.

The single-facility, multimodal experiments previously described exemplify the data management challenges associated with science applications of AI. Rapid alignment and preparation of multimodal, high-dimensional science data for AI remain an important challenge. In addition, experimental facilities, such as the Large Hadron Collider (LHC) and Relativistic Heavy Ion Collider (RHIC), are generating data at rates in the TB/s range for raw data with data volumes quickly approaching the exabyte scale[10]. Experiments using the High-Luminosity upgrade to the LHC will archive exabyte-scale datasets every year. DOE Leadership Computing facilities also run some of the largest-scale simulations for applications such as lattice quantum chromodynamics (LQCD), fusion energy, molecular dynamics (MD) of computational chemistry and biology, and direct numerical simulation (DNS) of computational fluid dynamics on combustion and climate. The volume of these simulations can be petabytes even without exascale supercomputers. The LHC experiments that discovered the Higgs boson in 2017 archived 200 PBs, which is equivalent to 3,000 years of ultra-high-definition video streaming. Deploying AI in these circumstances with extreme data volumes and rates poses challenges in establishing I/O data streams, staging and tiering storage, offering sustained computational performance, and providing accurate supporting metadata.

Controlling either extreme-scale simulation or experimental facilities with AI requires low-latency analysis and inference. Currently, most analyses of experimental and simulation data are done post hoc, after the experiment or simulation has run. To enable AI-driven automated control for faster scientific discovery, these high-volume, high-velocity data need to be analyzed in real time.

AI can aid science by generating hypotheses for questions such as: what kinds of experiments need to be done, or what kinds of questions need to be asked? Many science questions require a combinatorial approach, for example, to select and combine different raw materials to design a new functional material that has attractive properties or to identify biological conditions and genes to engineer and yield the best biofuel. The number of permutations to consider is well beyond what would be possible for humans or naive automation to explore.

The opportunity to “influence the development of AI tools by democratizing access to benchmark science data” (described in the next section) directly addresses these challenges.

[10] Albrecht, Johannes, et al. “A Roadmap for HEP Software and Computing R&D for the 2020s.” *Computing and Software for Big Science* 3.1 (2019): 7.

2.1.2 Science workflows with AI need human oversight

Roundtable participants described many inhibitors to efficient use of AI technologies in science applications, ranging from I/O capabilities for HPC systems to insufficient storage and access to data to the siloed software stacks for HPC, big data, and AI. Often, these challenges are overcome in ad hoc ways with bespoke solutions, requiring human oversight.

Science data are generated across a diverse set of facilities, instruments, and sensors that span a range of complexity and scales. In some cases, the data are too large to be stored and must be processed at the detector or streamed directly into a supercomputer for analysis in real time. In other cases, data generated at different facilities need to be integrated, which requires collocation of the datasets in common storage and computing resources.

The software used to process and wrangle data, build and deploy AI applications, and analyze results often are developed by distinct communities with little attention to facilitating combined workflows. Bridging these software silos is a major challenge.

The computation for a given research effort may take place on a variety of platforms, distributed geographically and with very different computing capacities (e.g., edge to HPC). There is a need for ensuring AI applications can be deployed seamlessly across platforms and on new and emerging architectures.

Per the roundtable participants, there is an opportunity to "make AI operational in science with composable services for simulation, data analysis, and AI at all scales."

2.1.3 There is no theory encompassing data, AI models, and tasks

Educated trial and error continues to guide advances in science applications of AI. Currently, there is no holistic theoretical approach unifying data, AI models, and the tasks performed by models that would help answer critical, foundational questions, such as: what information about a dataset can be deduced from a model trained on the data? Do models inherit the access limitations or classifications of the training data? For a given dataset and task, what are the best model, hyperparameters, and training method? When are more data needed, and how much incremental information will they have? Which data would make the biggest improvement? In what circumstances can a model be transferred to new data?

Presently, many science applications of AI begin with the desire to perform a particular task on a given dataset or data stream. Training data are chosen and prepared. Then, a model is identified and trained. The choice of model often involves some educated guess work. Furthermore, to improve the model's accuracy, it is not clear if more training data, a different model or hyperparameters, or the addition of domain knowledge would be beneficial. Currently, it is not possible to relate the models to previously collected data or determine what other tasks are relevant. The missing theoretical framework for data, AI models, and tasks has important implications for performance and optimization of AI in science, as well as overall productivity of AI applications and the reusability of models.

The lack of a unifying theory also influences decisions regarding the sharing of AI models and techniques. Because it is not known what attributes of a dataset can be derived from a model trained on that data, AI models trained on restricted data are not shared more broadly than the data themselves. Having a better understanding of the relationships between data and models would have an enormous impact in connecting research on restricted data with advances in open science.

The *Workshop Report on Basic Research Needs for Scientific Machine Learning* identified foundational PRDs in domain-aware, interpretable, and robust ML. These PRDs call for mathematical approaches for developing a foundational understanding of AI/ML, the lack of which is described here. In the next section, the opportunity to "address open questions in AI with frameworks for relating data, models, and tasks" describes a computational and data-driven approach for identifying key features of such a theory.

2.1.4 Applying FAIR principles to science data is challenging

AI research may have substantially different requirements for reusing data than other more domain-specific reuse cases. Roundtable participants showed unanimous support for FAIR principles but noted there are unique challenges in making science data FAIR for AI. Furthermore, there currently is much less attention on how to make AI models FAIR.

Scientific workflows can now involve several AI steps across different stages of the data life cycle, from data generation (from experiments or simulations) to data reduction (filtering, transformation, compression, etc.) to drawing conclusions from the data. Historically, when all the involved data could be tabulated on paper, scientific results were disseminated in written format. For example, a manuscript usually included details of the performed experiment, data acquired, and contributions based on the data, meaning that all of the required information, steps, and data to validate and replicate the experiment results would be found in the published paper. Now, data and models live outside the manuscript, and reproducibility of the results is contingent on the data and models being FAIR.

There are specific challenges associated with each of the FAIR principles when considering science data for AI:

Findability: How will the AI community search or browse for data? What attributes are important to include in the metadata that will further enable search and queries by AI researchers? The advancement of AI depends on large, well-characterized training datasets. Enabling researchers to find data requires an understanding of the relevant metadata that will be used for a search—either by a human or machine. We speculate that properties such as the structure, dimensionality, sparseness, and multimodality of the data, as well as information about the types of models trained on the data, will be, at least, as relevant as attributes such as discipline domain, source, author, and other information found in current metadata standards.

Some attributes relevant to search are contained within the data, and scalable queries across these resources require precomputing values across large amounts of raw or processed data.

These data need to flow between repositories to enable computations across the whole corpus of available data. Furthermore, the metadata standards and models have been established across a number of fields, but the application and enforcement of these standards are challenging, manual, and offer few incentives to curate metadata.

Accessibility: AI applications in science present new data access patterns, for example, training over federated data or distributed training and inference. AI applications also introduce different and unpredictable I/O patterns. In some cases, AI algorithms must be trained over geographically dispersed data repositories. This presents unique challenges when attempting to scale up access as certain AI algorithms require the storage system to read and reread entire datasets. This means that when a scientist wants to train a method on a large amount of data that may be stored in a High Performance Storage System (HPSS), these data must be restored to a file system or other storage that enables rapid and/or random access.

Interoperability: For data to be interoperable in the sense intended by the FAIR principles, a machine needs to be able to ingest and interpret data from different repositories. This implies that the data and metadata are described by vocabularies that follow FAIR principles. For instance, the metadata are linked to established ontologies to enable systematic linking between the datasets. Although there are many efforts underway to standardize metadata and create semantic links between datasets, this remains a difficult problem.

Beyond these challenges, there are fundamental open questions about how to use data from different sources in AI applications. Even sources producing the same type of data can introduce hidden biases and behave differently when presented to the same model. Some applications call for different types of data to be integrated, such as experimental and simulation data, and there is no principled way to do this currently.

Reusability: Machine readability of metadata, provenance, and annotations are essential for AI. Yet, the metadata needed for a given AI application can be difficult or impossible to know in advance. Metadata are critical for understanding biases in data and for interpreting results from AI applications. Like other types of data analysis, AI applications require detailed metadata, provenance, and annotation for interpretable, reliable, and transparent results.

Beyond FAIR data, there is value in extending the FAIR principles to AI models. However, there are even more challenges in making AI models FAIR. As for data, the challenges around findability of models are further complicated by insufficient study into how the AI R&D community searches and uses models. However, challenges involving accessibility, interoperability, and reusability reflect inherent open research questions about, for example, what can be inferred from a model about the underlying dataset, or if the model should inherit access limitations of the training set; transfer learning; explainability; the scarcity of standard structures and schemas for AI models; and the lack of a unifying framework for data, models, and tasks.

Currently, there are several different types of formats for storing models, and there are compatibility issues that arise among them. In the case of traditional ML, model storage may be interpreted as the storage of the source algorithm and required input parameters. For DL, storage will require storing parameters (e.g., weights), network architecture, and dynamic execution graphs. The weights are driven by the training data, and there is a question of how much training data need to be stored with the model for validation and reproducibility. There have been efforts to standardize DL model formats. Two of the most promising model exchange formats are NNEF (Neural Network Exchange Format) and ONNX (Open Neural Network eXchange), which is supported by major DL frameworks, such as PyTorch, CNTK, MXNet, and TensorFlow. The current model exchange format is not, however, linked back to the data used in training.

Addressing this challenge will require collaborations between domain scientists, AI experts, and data management experts to better understand the needs of the AI community, as well as to refine what is needed to reuse science data as data for AI. Such collaborations could form around, for example, science benchmark data or the creation of frameworks for data and AI models (detailed in the next section).

2.2 Opportunities

This section describes opportunities for the DOE SC to advance AI R&D and improve the impact of AI tools for the SC mission. Each opportunity addresses one or more of the aforementioned research challenges.

2.2.1 Influence the development of AI tools by democratizing access to benchmark science data

Making science data available to AI researchers and developers will improve the utility and performance of AI tools for science. Benchmark datasets are used to compare analysis, AI, or other computational methods. It has been argued that the ImageNet[11] competition and benchmark dataset sparked the DL revolution, specifically by demonstrating the capabilities of convolutional neural networks in object recognition tasks in natural images. Similarly, the advancement of AI research for DOE SC applications starts with facilitating access to science data. This opportunity directly addresses the challenges outlined in Section 2.1.1 and 2.1.4.

Envisioned here are published benchmark datasets that exemplify the distinguishing attributes of science data (Section 2.1.1) with appropriate storage, access rights, and integration with computational capabilities and analysis tools to focus the development of AI tools and techniques on science needs. Benchmark datasets will have greater impact if domain experts help define the data and metadata and provide explanations about what phenomena they describe and the nature of tasks that can be performed on them. Challenges, citizen science competitions, and partnerships can provide a formal context and focus for developing the necessary metadata and documentation and also attract new talent to the SC mission.

[11] www.image-net.org

Metadata, provenance, and annotations are key components for making data FAIR for AI. These can be improved for a benchmark dataset through coordination and feedback from the AI community. A feedback loop of discoveries, annotation, and updates to metadata should make it possible to associate learned information with benchmark datasets. One way to facilitate this feedback loop is to ensure that systems provide full visibility into the workflow and data provenance characteristics of the datasets.

2.2.2 Make AI operational in science with composable services for simulation, data analysis, and AI at all scales

Composable services can enable the efficient execution of scientific workflows of simulation, data analysis, and AI across the computing continuum, from edge to HPC. The vision is for combined infrastructure and software capabilities that reduce data movement and analysis at all scales; federate data and computing resources for seamless AI workflows, incorporating data collection, edge computing, and AI; optimize data placement and organization in storage and memory hierarchies to reduce data movement and associated processing latencies; and integrate heterogeneous computing architectures and new hardware.

To support the data flow from geographically dispersed sources, the networking and software infrastructure must, in some cases, perform computations where the data are stored, and, in others, move the data to where the computations can be performed. This may require federated access to resources at different facilities so that, for example, analyses and data transfers between facilities can be initiated programmatically without human intervention. Intelligent infrastructure design is needed to ensure that the data movement (communication) is minimized to avoid bottlenecks. Data movement services that span facilities, supercomputer memory hierarchies, file systems, and tape archives will enable scientists to focus their efforts on the analysis as opposed to data management. In certain AI algorithms, data are read and reread many times to train models. Therefore, reducing latencies in data access can diminish the training time for ML techniques dramatically.

To meet increased demands for computing, hardware accelerators are needed to reduce the overall power footprint. Scientists need access to new hardware tailored for different AI applications as it becomes available in order to adapt or develop algorithms optimized for the new architectures. Access to cutting-edge testbeds in the DOE ecosystems can be enabled through a common application program interface (API) and federated identity management across the DOE SC.

There is an opportunity to advance AI research with appropriate computing resources to run and cross validate models on different datasets without having to worry about dependencies or specific hardware architecture tuning. Container[12] technology is helpful in this regard, but it is not readily usable by non-experts because the technology itself continues to evolve. HPC facilities could maintain sets of optimized containers for the various architectures they support.

[12] A container is a way to encapsulate software and library dependencies into a single package that can run on a variety of systems, requiring relatively minimal underlying system configurations.

Research is needed to make containers easier to build and port across cloud and HPC resources. In addition, published datasets and models need to be easily accessible through extensions to the data science software stack in the same way image recognition benchmark datasets can be directly imported into a DL workflow.

2.2.3 Address open questions in AI with frameworks for relating data, models, and tasks

Frameworks for tracking relationships among data, models, and tasks can address strategically important open questions in AI research, such as those highlighted in Section 2.1.3 and the challenge of making data findable for the AI R&D community. Envisioned here is a framework that would link all salient aspects of an AI workflow, including the data, AI model, task, training methodology, and accuracy metrics and measures. An important feature of this framework is the holistic view of this workflow.

With the accumulation of many such workflows in the framework, researchers will be able to discover higher-level patterns among the workflows that reveal a deeper understanding of the relationships between data, models, and tasks. Some of these higher-level patterns are already known. For example, in the context of DL, it is known that convolutional neural networks currently work best as a model for image-based tasks (image recognition, segmentation, deblurring). Finding similar connections between other science-relevant data modalities and abstract models is an open research question. There also are examples where transfer learning (where a model trained for a specific task or dataset is used for a different task with minimal to no modifications) works well, and examples where it has failed. The framework envisioned here could help identify reasons for these successes and failures. The framework would inform investigation on topics, such as active learning, AutoML, and transfer and lifelong learning.

Some current large-scale services enabling the findability and accessibility of data and models include OSTI's DataID Service[13], DLHub[14], OpenML[15], and Zenodo[16]. Unique and persistent identifiers, such as digital object identifiers (DOIs), are central to these efforts.

Relationships among data, models, and tasks could be efficiently captured at the point of publication by including these elements as part of the scholarly record. Almost all data science development platforms and languages now offer "notebooks," an interactive interface to interleave analysis code with formatted text, figures, and equations. Jupyter notebooks are arguably the most popular medium for AI practitioners to develop and share AI models and the models' provenance, including various forms of different model cross-validation techniques and hyperparameter tuning recipes (automatic or manual). MATLAB and R also offer similar features

[13] https://www.osti.gov/data-services
[14] Chard, Ryan, et al. "DLHub: Model and Data Serving for Science." *arXiv preprint arXiv:1811.11213* (2018).
[15] Vanschoren, Joaquin, et al. "OpenML: networked science in machine learning." *ACM SIGKDD Explorations Newsletter* 15.2 (2014): 49-60.
[16] https://zenodo.org/record/2541184#.XSOLvOhKju0

through MATLAB live editor[17] and R Notebooks[18], respectively. There also is support for running these notebooks on various Leadership Computing Facility resources (JupyterHub) and cloud computing services (Google Colab). Publishing a scientific manuscript with its accompanying dataset and the analysis notebook that includes all the data transformations and visualizations could help to capture the relationships among data, models, and tasks.

2.3 Enabling Capabilities

This section presents a number of prerequisite, enabling capabilities for addressing the aforementioned opportunities, as well as a broad range of data science and data management options. Like the opportunities, these enabling capabilities identify areas where additional DOE SC investments would be impactful. However, the opportunities described in Section 2.2 directly address the challenges identified in Section 2.1. By contrast, these enabling capabilities are more foundational and would impact data science more broadly than the AI-focused opportunities.

2.3.1 Data management support and incentives for teams generating data

There is a need to support domain science teams in the production of FAIR data by linking them with data science and data management experts and providing incentives for data management. Improving access to expertise in data science and data management, AI best practices, metadata standards and ontologies, and data sharing and retention opportunities can help research teams make their data FAIR and ready for AI. Researcher engagement with AI experts, research libraries, archives, and community organizations, such as the Research Data Alliance,[19] can increase capabilities and ensure alignment between best methods, community standards, and DOE research needs. The engagements and application of the FAIR principles should run from experimental design through to final data publication.

[17] https://www.mathworks.com/products/matlab/live-editor.html
[18] https://bookdown.org/yihui/rmarkdown/notebook.html
[19] https://www.rd-alliance.org/

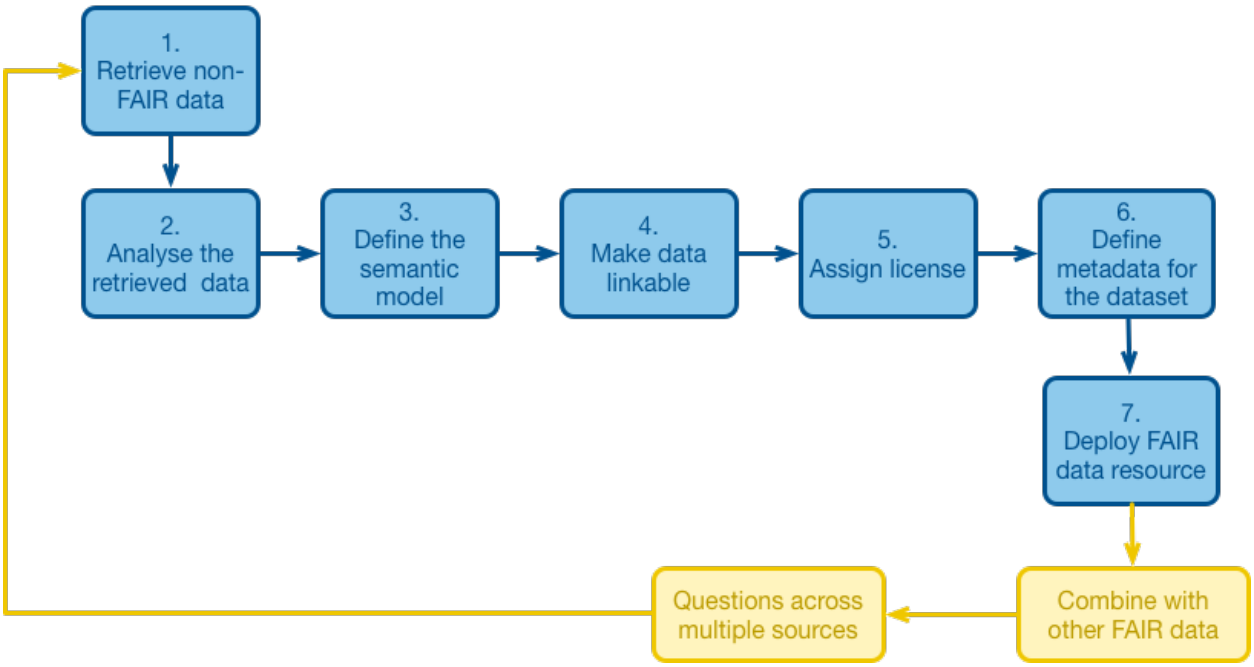

Figure 3: Diagram showing the: "FAIR-ification of Data" (https://www.go-fair.org/fair-principles/fairification-process/).

Figure 3 shows a generic pipeline for making data FAIR with the possibility of linking FAIR datasets from different sources (yellow boxes). For SC-relevant applications, Step 3 is particularly challenging and requires both deep domain knowledge of the data and expertise in best practices for metadata and ontologies. In some cases, ontologies already exist and can be adopted. However, in many DOE scientific domains, such ontologies need to be defined and built from the ground up. Replacing semantic models with domain-specific forward models is an alternative approach. For instance, in X-ray diffraction imaging experiments, the underlying physics forward model of how a diffraction pattern image from a sample is measured on a pixelated detector is known to experimentalists in this field. Therefore, the semantic model parameters of a diffraction dataset are defined as the forward model constants that led to the data generation process. In this specific case, they can include parameters like X-ray beam energy used, scanning parameters, optics and beamline configurations, detector and sample properties, and/or noise models due to various sources of error. These parameters aid in defining the dataset metadata (Step 6), which, in turn, enable combining and linking FAIR datasets together.

Roundtable participants envisioned data scientists and data management experts to be embedded within domain teams. Embedding should begin as early in the process as possible, from experimental design through the final analysis. Domain experts bring years of experience and understanding of data, metadata, and provenance. On the other hand, domain scientists may find it challenging to adopt the mindset of non-specialists or fail to document deeply embedded assumptions about the data. A close collaboration between domain and data experts can capture the best of both worlds: domain expertise and robust, scalable pipelines to produce high-value FAIR data.

Such relationships also will play a role in educating domain scientists about AI and data management. As the power of AI techniques progresses, domain scientists in the DOE must become data and AI conversant. Processes and incentives to build a data-savvy workforce are important pieces in the overall puzzle. Given the challenge of hiring and retaining data scientists, modernizing the skill sets of on-staff domain scientists across SC is critical, including retraining programs.

Roundtable participants recognized challenges in attracting and retaining data science and data management expertise into partnerships with science domains. For example, publication cycles in some domain science fields are much longer than in data science and data management. It is important to support mechanisms that allow for collaborations between domain scientists and data scientists to publish on the timescale that is natural for each of their fields. Offering authorship to data scientists on domain science papers is helpful but not enough to drive recognition. New career paths or re-imaginings of existing paths (e.g., data librarians, data curators, etc.) may also help. These roles increase capacity for important expertise that may not fit in the current scheme for career advancement. Finally, hiring and retention are complex problems, especially given salaries for equivalent work are higher in industry. It is important to emphasize the SC mission, impact, and freedom of inquiry to attract a capable workforce. DOE SC also may choose to embrace its role as an early-stage developer of data science and data management talent for industry.

One opportunity in this space is finding a way to recognize members of the scientific community for contributions to a repository (either through curation or data contributions), perhaps with some formal incorporation into their impact score. The recognition could include indicators for quality and FAIR-ness. Such recognition could be used by funding agencies as part of funding decisions or by research institutions as part of promotion decisions. Data repositories could help report and share information about contributions made to the repository. This could be supported further through measures of the number of downloads or uses of the data with attribution awarded to the data contributor. Other incentives can be provided in the form of additional storage space or bandwidth to the system to facilitate access.
Community data repositories play a strategic role as keepers of domain-specific ontologies and standards and can provide incentives for data submitters to adhere to quality standards. Clearer and more detailed expectations from funding agencies and journals with respect to data management also can help incentivize best practices and maintain alignment with researcher career goals.

2.3.2 Automated collection of metadata, provenance, and annotations at scale

There is a need to reduce researcher burden and improve the quality of data with the automated collection of metadata, provenance, and annotations at scale. Machine-readable metadata, provenance, and annotations with standards would dramatically increase the FAIR-ness of data for AI and other analyses. The ability to automatically collect this information at scale so that data and metadata are “born digital” (rather than imported from paper copies) can reduce the burden on researchers and improve the quality of this information. Automated collection of metadata, provenance, and annotations are needed because the scale (volume and velocity) of

data often outstretch human capabilities and AI, in turn, is introducing autonomous decisions and transformations into science workflows.

Acquisition, curation, and storage of data frequently are the dominant bottlenecks for reusing science data for AI. Data producers encounter an array of challenges, including decisions regarding which data to acquire, which data to save, how to record metadata, and how to preserve provenance of the data pipeline. These are particularly challenging problem areas because science datasets tend to be very large (and so require specialized acquisition, storage, and access technologies), complex (and may not fit in many popular tabular formats), and highly distributed (many different labs and principal investigators [PIs] may contribute data relevant to any given scientific question). Data management is further complicated by the need for appropriate long-term retention policies (i.e., some datasets may be truly irreplaceable but expensive to store in full) and for access policies that can support a wide variety of users with the appropriate level of security for the storage facility. With an end goal of creating FAIR data at scale, there is a need for "smart" data collection and processing infrastructure that can automate the entire data pipeline with on-the-fly compression, cleaning, alignment, and annotation.

The need for automated data pipelines starts during acquisition itself. High-fidelity simulations and high-velocity experiments produce extreme quantities of data, straining retention policies and requiring *in situ* data analysis with irreversible compression or rejection. For the most extreme cases, ML algorithms are needed at the edge, i.e., at the sensor or simulation node itself. *In situ* analysis also can guide science "in the loop," adapting experiments to measure the highest value parameters or choosing (and even recommending) to run the most critical simulations. Finally, data acquisition should be designed with ML applications in mind. For instance, multimodal measurements should be captured at the detector and labeled with time stamps to permit correlation analyses. As data rates grow, automation-for-AI and AI-driven automation will become intrinsic to the data collection process.

To preserve the dataset's value after acquisition, producers should record detailed metadata of methods, conditions, and manipulations. Complete metadata enables analysis by subsequent data consumers who did not participate in the acquisition. Metadata should capture details of compression or rejection during acquisition (which might bias data), data processing (backgrounds, noise, systematics), and experimental or computational parameters. Where ML models are used in the data pipeline (e.g., denoising), the models themselves should be captured as metadata. A detailed history of sample preparation empowers data reuse and enables applications in AI algorithm development.

Even after a dataset is recorded or published, user interaction has value to downstream data consumers. At present, no framework is widely accepted for capturing usage as metadata. Opportunities here include recording annotations of the data acquisition (e.g., logbooks), data "likes" from previous analysis (i.e., popularity of subsets of the data in prior analysis), statistics of data interactions, and archiving of ML models used throughout the analysis. Today, human-

data interface typically is one way, displaying data for the user. An opportunity exists in the other direction by capturing the users' interactions in the dataset.

Part of the challenge for data provenance is cultural with a need to incentivize data producers to maintain value for unrelated data consumers (e.g., through data citations). Automation also can play a role in promoting FAIR data by lowering barriers to capturing provenance, metadata, and annotations.

SC-supported research communities are producing science data at ever-increasing rates. Information, data, and metadata that are born digital can help minimize human effort needed to make data FAIR. For instance, each photo taken with an iPhone has certain associated metadata: GPS, time of day, phone model, user account, resolution—more than 460 tags. These tags make it possible to determine whether or not an image has been altered (reusable), if the format is compatible with other devices (interoperable), and also power many of the image recognition algorithms. This is all made possible because individuals do not need to enter these tags manually: they are automatically captured by the device. The roundtable participants envisioned something analogous for science sample collection that captures GPS, date/time, PI, and other data that could be automatically assigned when the samples are logged. Today, automatic tagging with a desired level of flexibility for AI is unavailable in a large class of machines. While large detector complexes (such as in high energy physics and nuclear physics) collect elaborate metadata, directing this toward AI purposes is a required capability. Data scientists and domain scientists can and will re-annotate and re-tag data with new information to improve their discovery pipelines, but the initial context provided by data that are born digital accelerates this process. The provenance and lineage of the data also become a part of the additional information associated with the metadata for a dataset, facilitating reusability, verification and validation, and trust.

2.3.3 Scalable human interfaces for data

Researchers need enhanced capabilities for extracting information from data through scalable human interfaces for data. Tools and frameworks are needed to help data users find, understand, and reuse data. There is an opportunity to go beyond keyword searches and hit lists to visual interfaces for data and their relationships so that missing information or corroborations among research findings can be easily identified. This interface should consider other research products, including models, code, and publications. There is an opportunity to search and discover data based on new attributes important to AI research, which may not be captured by current metadata standards that address discipline domain, source, author, etc., or for the interface to suggest new directions of inquiry or new datasets to explore.

Data standards are critical to enabling interoperability among data and to facilitate the interface envisioned here. Standards, however, must be designed carefully to avoid inhibiting systems by narrowly building to a limited set of syntax (formats) or semantics.

As the amount of available data continues to grow, scalable technologies to find and retrieve datasets will revolutionize AI and allow scientists to harness the AI innovations to apply to the

datasets in their field. Exposing preexisting datasets into this interface will require revisiting, scanning, and re-annotating data, especially those datasets not born digital. Formatting and structuring such datasets in an automatic manner, albeit with human oversight, remain a continuing challenge.

Once relationships among datasets are established, there is an additional challenge in creating useful human interfaces and recommender systems. Such developments would benefit from a deeper understanding of how research communities interact with data and reason about the information they contain.

2.3.4 Strategic approaches to managing data management costs and resources

As data volumes increase, strategic approaches to managing cost and resources with respect to storage, data preparation, and curation are needed. These will depend on evaluating potential impact from data as a way to guide investments and support for curation and preservation, as well as exploiting new technologies and economies of scale, particularly with respect to storage. Although the cost of data storage has decreased rapidly over the past two decades, it has been outpaced by the demand for storage due to the growth of science data. This will force the science community to come up with judicious cost models and approaches to provide scalable storage and associated data handling capabilities for data for AI.

An important driver of investments will be the evaluation of the utility of the datasets. Valuable datasets, such as those collected from non-replicable experiments (e.g., observational astronomy), will need to be stored in as pristine a manner as possible. On the other hand, simulation-generated datasets may not be preserved if the cost of rerunning the simulation is less than the cost of storing the dataset for the long term. These decisions and trade-offs will need to span the data life cycle.

Quantifying the value of a dataset is an open research question. A sound methodology for defining dataset quality is in its early stages and, ideally, requires an information-theoretic understanding of how the collected data impacts AI model selection. Data value should include principled ways of understanding what data are collected, what is worth retaining, and when it can be reproduced or regenerated. The value may vary according to the domain use and the particular models and tasks that employ the data. There are many potentially conflicting metrics for value, including the potential cost to reproduce the data and frequency of use. Certain datasets can appear poor in quality in isolation but prove invaluable when combined with other datasets. The value of a dataset also may grow in the future as new analysis methods or related datasets emerge.

Assessing the value of datasets will necessarily foster questions of ownership and responsibility. There is an opportunity for the DOE SC to engage scientific communities in discussions concerning these issues. Guidance from SC, informed by these discussions, would help to reduce ambiguities around roles and responsibilities for data retention and curation.

3. About the Roundtable

This one-day roundtable included 35 experts from 12 DOE national labs, NIH, and NSF. Participants had wide-ranging expertise in areas such as AI/ML, data management, data curation, metadata, library sciences, storage systems and I/O, open data, big data, and edge computing. These experts represented mission drivers across the six DOE SC programs and OSTI with ties to Office of Science-supported research activities, science user facilities, and community data repositories.

The discussions proceeded in four phases, indicated in the agenda (Figure 4). During the first part of the morning, participants presented lightning talks on using research data for AI/ML in science. Presenters were encouraged to share "success stories," as well as "frustration stories." Before and after lunch, there were parallel breakout sessions. The sessions before lunch had predetermined themes around the idea of making data FAIR (Findable, Accessible, Interoperable, Reusable)[20] for AI. The afternoon breakouts focused on topics that emerged from the morning sessions: Storage and Data Placement at all Scales, the Role of the Data Scientist, Metadata, and Making Data and Models FAIR Together. Finally, the day concluded with plenary readouts from the breakout sessions and a discussion about what was learned, potential synergies among the ideas presented, and potential gaps. Common cross-cutting themes that emerged included: Interoperability of Data from Different Facilities/Data Sources, the Need to Better Understand the Data Landscape, and the Need to Understand and Assess the Value of Data.

Discussions in the breakouts and final plenary session were highly interactive. Participants were encouraged to form small groups for brainstorming and exploring ideas in depth. The breakouts had facilitators who set expectations, guided conversations, and kept the participants on schedule. The discussions were mediated with the help of sticky notes, white boards, flip charts, and digital media. Organizers and scribes took care to record all of the conversation artifacts.

After the roundtable, a writing team of six experts from across the DOE SC labs synthesized the materials collected through the day into the findings and framework presented in this report. A summary of these findings was first presented at a meeting of the Advanced Scientific Computing Advisory Committee on September 23, 2019.[21]

[20] Wilkinson MD et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018 DOI: 10.1038/sdata.2016.18

[21] Presentation slides from the September, 2019 ASCAC meeting can be found here: https://science.osti.gov/ascr/ascac/Meetings/201909

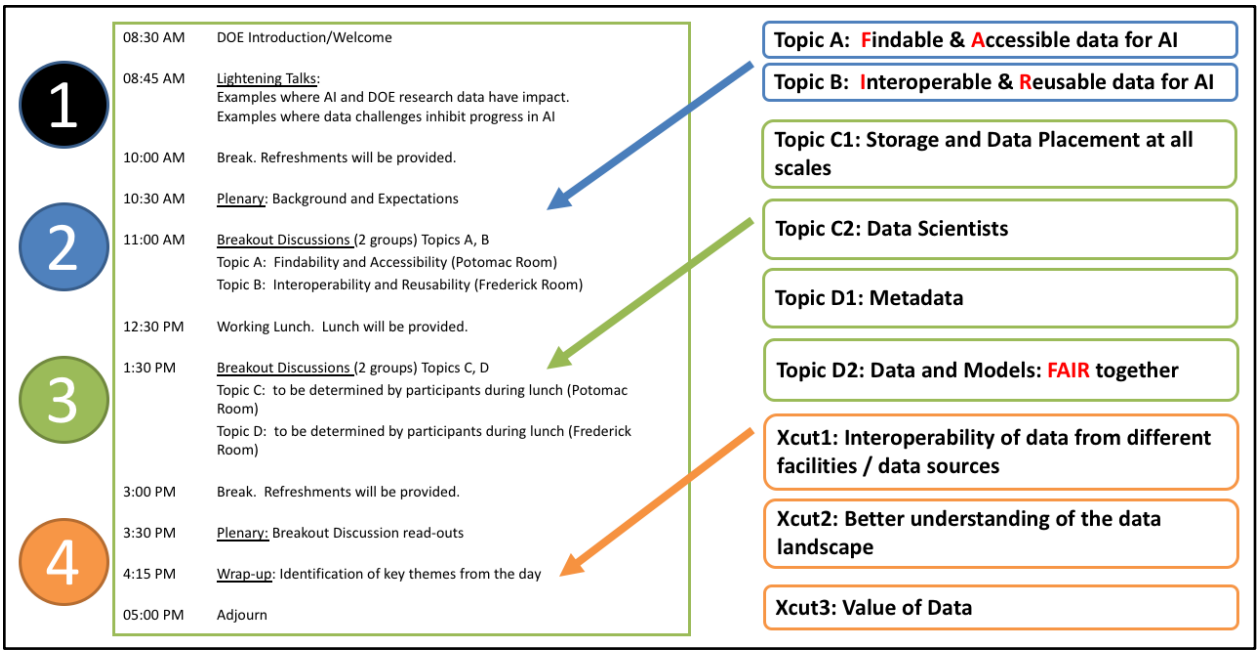

Figure 4: Office of Science Roundtable on Data for AI Agenda

4. Conclusions

The fundamental finding of this roundtable is that there are opportunities to advance AI R&D and increase the benefit of AI to science by improving the reusability of science data and AI models and through the development of methodologies and services to integrate AI seamlessly and routinely into science workflows. The roundtable participants identified three priority opportunities for data to advance AI in science:

1) Influence the development of AI tools by democratizing access to benchmark science data
2) Make AI operational in science with composable services for simulation, data analysis, and AI at all scales
3) Address open questions in AI with frameworks for relating data, models, and tasks.

These opportunities are presented in a broader context of open research challenges in AI and prerequisite, enabling capabilities in data science and data management.

5. Appendix A: Glossary

AI	Artificial Intelligence. In this report, we consider AI to be inclusive of machine learning (ML), deep learning (DL), neural networks (NN), computer vision, and natural language processing (NLP).
AI model	An AI model is an inference method that can be used to perform a "task," such as prediction, diagnosis, classification, etc. The model is developed using training data or other knowledge.
AI task	The inference activity performed by an artificially intelligent system.
AI tools	AI tools, such as PyTorch and TensorFlow, used to build and deploy AI applications.
Active learning	A research field focused on data-efficient machine learning algorithms that are able to query the dataset or data source for new training samples.
AutoML	Stands for automated machine learning, not to be confused with a Google toolkit with the same name. It also is the process of automatically finding the model and model hyperparameters that best describe a particular training dataset.
Composable	Composable services are created from interoperable modular components that can be assembled flexibly into multiple well-defined functional and usable tools or capabilities.
Data for AI	The digital artifacts used to generate AI models and/or used in combination with AI models during inference.
DL	Deep Learning
Lifelong learning	Also continuous learning. A strategy for dealing with a well-known shortcoming of artificial neural network approaches, namely catastrophic forgetting, where the model's performance degrades on previously learned tasks as new tasks are introduced.
ML	Machine Learning
Ontology	The models of knowledge and associated definitions and relationships among terms or categories that are essential for interoperability among datasets.
Transfer learning	The act of using pre-trained models for tasks/data other than what the models were originally designed for.

Cover design by Y. Belyavina, Brookhaven National Laboratory (2019).

U.S. DEPARTMENT OF ENERGY | Office of Science