U.S. DEPARTMENT OF ENERGY | Office of Science

Office of Science
Data for AI Round Table

ASCAC

September 23, 2019

Laura Biven, PhD
Program Manager, Computer Science,
Advanced Scientific Computing Research
U.S. Department of Energy
Laura.Biven@science.doe.gov

DOE is uniquely positioned to lead in AI/ML R&D

DOE has a unique combination of capabilities and infrastructure to lead the Nation in Artificial Intelligence (AI) and Machine Learning (ML) research and development:

- A broad **mission** that presents new and unique research problems on national and global scales 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** 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 scientists, computer scientists, and mathematicians currently engaged in AI and related fields.

AI Initiative

Enhance access to high-quality and fully traceable Federal data, models, and computing resources to increase the value of such resources for AI R&D, while maintaining safety, security, privacy, and confidentiality protections consistent with applicable laws and policies.

Executive Order 13859 of February 11, 2019
Maintaining American Leadership in Artificial Intelligence

Round Table Focus, Goals, & Context

SC Round Table on Data for AI

Focus:

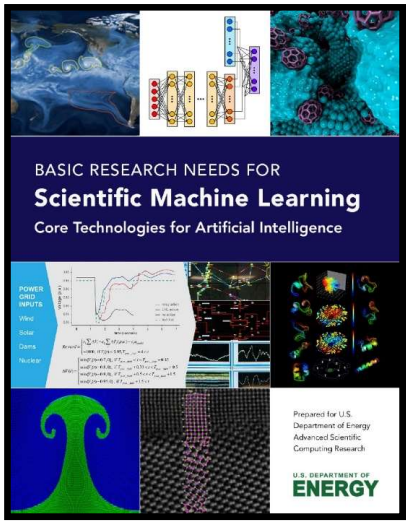
Enhancing and enabling access to high-quality and fully traceable research data, models, and computing resources to increase the value of such resources for AI R&D and the SC mission.

Goal:

Identify key challenges/opportunities and potential next steps for the Office of Science.

Complementary

Synergistic



AI FOR SCIENCE TOWN HALL

Other AI-focused workshops from SC program offices

Office of Science (SC) Round Table on Data for AI

Rockville Hilton

June 5, 2019

https://www.orau.gov/RTD-AI2019/

Supported by the **Office of Science Working Group on Digital Data** (SCWGDD), which includes representatives from all six SC program offices and OSTI

POC: Laura Biven (ASCR)

Participants came from 12 DOE National Labs, NIH, & NSF with expertise in AI/ML, data management, data curation, metadata, library sciences, storage systems and I/O, open data, big data, and edge computing; with ties to SC research, facilities, and community data repositories

Agenda & Breakout Topics

1

08:30 AM	DOE Introduction/Welcome
08:45 AM	Lightening Talks: Examples where AI and DOE research data have impact. Examples where data challenges inhibit progress in AI
10:00 AM	Break. Refreshments will be provided.

2

10:30 AM	Plenary: Background and Expectations
11:00 AM	Breakout Discussions (2 groups) Topics A, B Topic A: Findability and Accessibility (Potomac Room) Topic B: Interoperability and Reusability (Frederick Room)
12:30 PM	Working Lunch. Lunch will be provided.

3

1:30 PM	Breakout Discussions (2 groups) Topics C, D Topic C: to be determined by participants during lunch (Potomac Room) Topic D: to be determined by participants during lunch (Frederick Room)
3:00 PM	Break. Refreshments will be provided.

4

3:30 PM	Plenary: Breakout Discussion read-outs
4:15 PM	Wrap-up: Identification of key themes from the day
05:00 PM	Adjourn

Topic A: Findable & Accessible data for AI

Topic B: Interoperable & Reusable data for AI

Topic C1: Storage and Data Placement at all scales

Topic C2: Data Scientists

Topic D1: Metadata

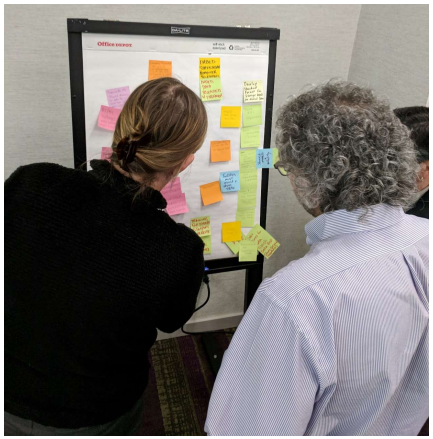
Topic D2: Data and Models: FAIR together

Xcut1: Interoperability of data from different facilities / data sources

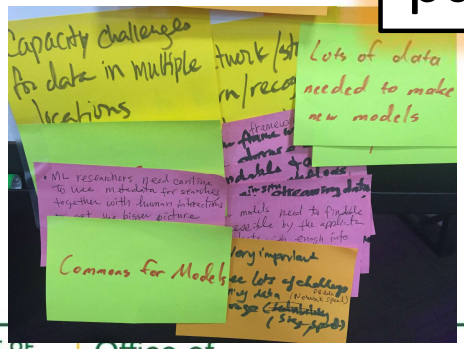
Xcut2: Better understanding of the data landscape

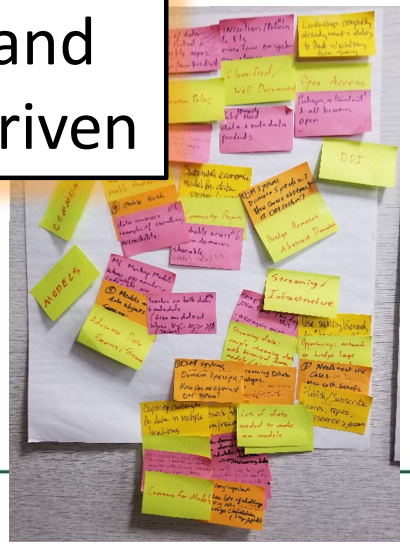
Xcut3: Value of Data

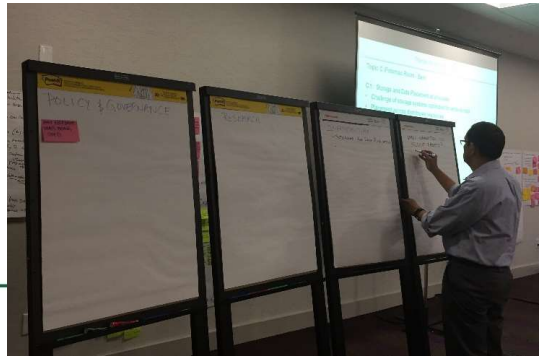
Process



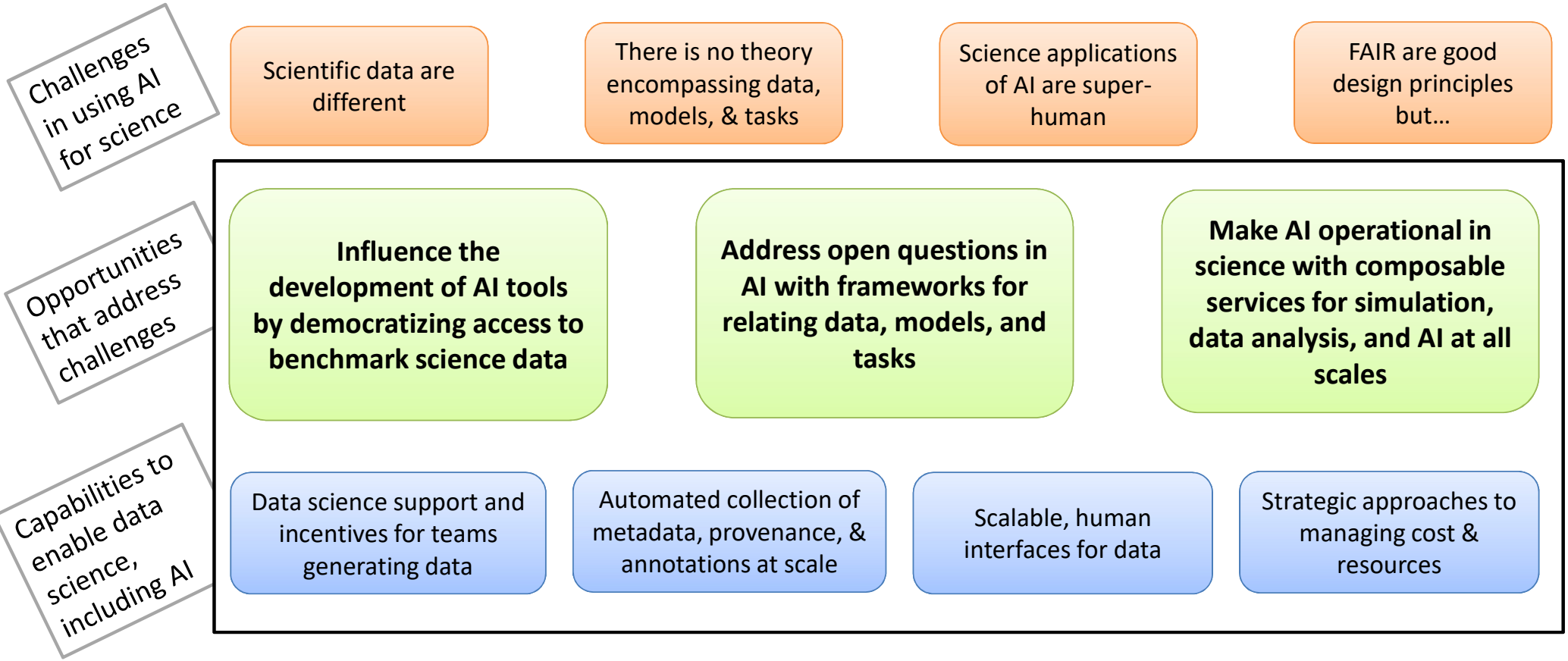
Discussions were interactive and participant-driven







Office of Science Data for AI Round Table: Challenges, Opportunities, & Enabling Capabilities



Thank you!!

SC Organizing Team

SC Working Group on Digital Data

POC: Laura Biven (ASCR)

Ben Brown (ASCR)
Michael Cooke (HEP)

Mariam Elsayed (SC-2)
Sujata Emani (AAAS - BER)
Jay Hnilo (BER)
Carolyn Lauzon (ASCR / AI Office)
Joanna Martin (OSTI)
Jessica Moerman (AAAS - BER)

Lab Writing Team

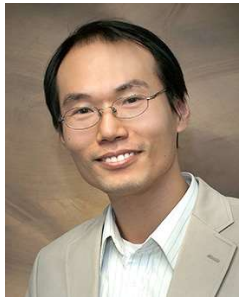
Kjiersten Fagnan (LBNL/JGI)

Youssef Nashed (ANL)

Gabriel Perdue (FNAL)

Daniel Ratner (SLAC)

Arjun Shankar (ORNL)



Shinjae Yoo (BNL)

BACK UP

Challenges

Scientific data are different

There is no theory encompassing data, models, & tasks

Science applications of AI are super-human

FAIR are good design principles but…

- Science data are different from the data typically available to AI research communities.
- Scientific data can be high-dimension, multi-modal, complex, sparse,…

Challenges

Scientific data are different

There is no theory encompassing data, models, & tasks

Science applications of AI are super-human

FAIR are good design principles but…

- Open research questions include:
  - 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 data?
  - For a given dataset and task, what are the best models, hyper-parameters, and training methods?
  - When is more data needed? How much incremental information will it have? Which data would make the biggest improvement?
  - In what circumstances can a model be transferred to new data?

Challenges

Scientific data are different

There is no theory encompassing data, models, & tasks

Science applications of AI are super-human

FAIR are good design principles but…

- For science, the goal is often not to recreate human intelligence as it is for AI research, but to have super-human capabilities in data analysis, control, hypothesis generation,…
- Science applications where ML/AI is impactful are characterized by extreme data rates and volumes, high complexity, extremely low latency requirements, …
- Scientific questions often require the combination of data from multiple sources, including data from across the SC facilities.

Challenges

Scientific data are different

There is no theory encompassing data, models, & tasks

Science applications of AI are super-human

FAIR are good design principles but…

- …, for AI reuse, there are open questions:
  - **Findability**: How will the AI community search or browse for data? What attributes are important to include in the 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 present different and unpredictable I/O patterns.
  - **Interoperability**: There are open questions about how to use data from different sources in AI applications.
  - **Reusability**: The metadata needed for a given AI application can be difficult or impossible to know in advance. Machine-readability of metadata, provenance, and annotations are more important for AI.

Opportunities

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

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

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

Making science data available to AI researchers and developers will improve the utility and performance of AI tools for science

- Benchmark science datasets that reflect the unique attributes of scientific data can focus the development of AI tools and techniques on science needs.
- Benchmark science datasets can also focus development on impactful applications, for example, facility operations and control.
- Challenges, citizen science competitions, and partnership around these can enhance the value of benchmark data.

Opportunities

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

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

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

Frameworks for tracking relationships between data, models, and tasks can address strategically important open questions in AI research

- Access to data, models, tasks, and their relationships can encourage an "empirical" approach to addressing open questions in AI research.
- 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.
- Tracking relationships among data, models, and tasks can also improve the reproducibility of AI research.

Opportunities

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

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

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 science workflows of simulation, data analysis, and AI across the computing continuum from edge, to HPC.

These combined infrastructure and software capabilities must:

- reduce data movement and analysis bottlenecks at all scales
- federate data and computing resources for seamless AI workflows, incorporating data collection, edge computing, and HPC.
- optimize data placement and organization in storage and memory hierarchies to reduce data movement and associated processing latencies
- incorporate heterogeneous computing architectures and new hardware

Enabling Capabilities

Data science support and incentives for teams generating data

Automated collection of metadata, provenance, & annotations at scale

Scalable, human interfaces for data

Strategic approaches to managing cost & resources

Improving access to expertise in data science/AI best practices, metadata standards and ontologies, data sharing and retention opportunities can help research teams to make their data FAIR.

- Researcher engagement with AI experts, research libraries, archives, and community organizations can increase capabilities and ensure alignment between best methods, community standards, and DOE research needs. Engagement should run from experimental design through to final data publication.

Community data repositories act 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.

Enabling Capabilities

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

Machine readable, "born-digital" 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 can reduce burden and improve quality.

Enabling Capabilities

Data science support and incentives for teams generating data

Automated collection of metadata, provenance, & annotations at scale

Scalable, human interfaces for data

Strategic approaches to managing cost & resources

Tools and frameworks are needed to help data users find, understand, and reuse data

- There is an opportunity to go beyond key word searches and hit lists to a cartography of data and their relationships to help identify missing information or corroborations among research findings and, ultimately, to understand the system-of-systems represented by the. This interface could incorporate 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 topic, format, etc.

Enabling Capabilities

Data science support and incentives for teams generating data

Automated collection of metadata, provenance, & annotations at scale

Scalable, human interfaces for data

Strategic approaches to managing cost & resources

As data volumes increase, strategic approaches to managing cost and resources are needed.

- These will depend on:
  - Evaluating potential impact from data as a way to guide investments in and support for curation and preservation
  - Exploiting new technologies and economies of scale, particularly with respect to storage

FAIR Data Principles

To be Findable:

- F1. (meta)data are assigned a globally unique and persistent identifier
- F2. data are described with rich metadata (defined by R1 below)
- F3. metadata clearly and explicitly include the identifier of the data it describes
- F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:

- A1. (meta)data are retrievable by their identifier using a standardized communications protocol
- A1.1 the protocol is open, free, and universally implementable
- A1.2 the protocol allows for an authentication and authorization procedure, where necessary
- A2. metadata are accessible, even when the data are no longer available

To be Interoperable:

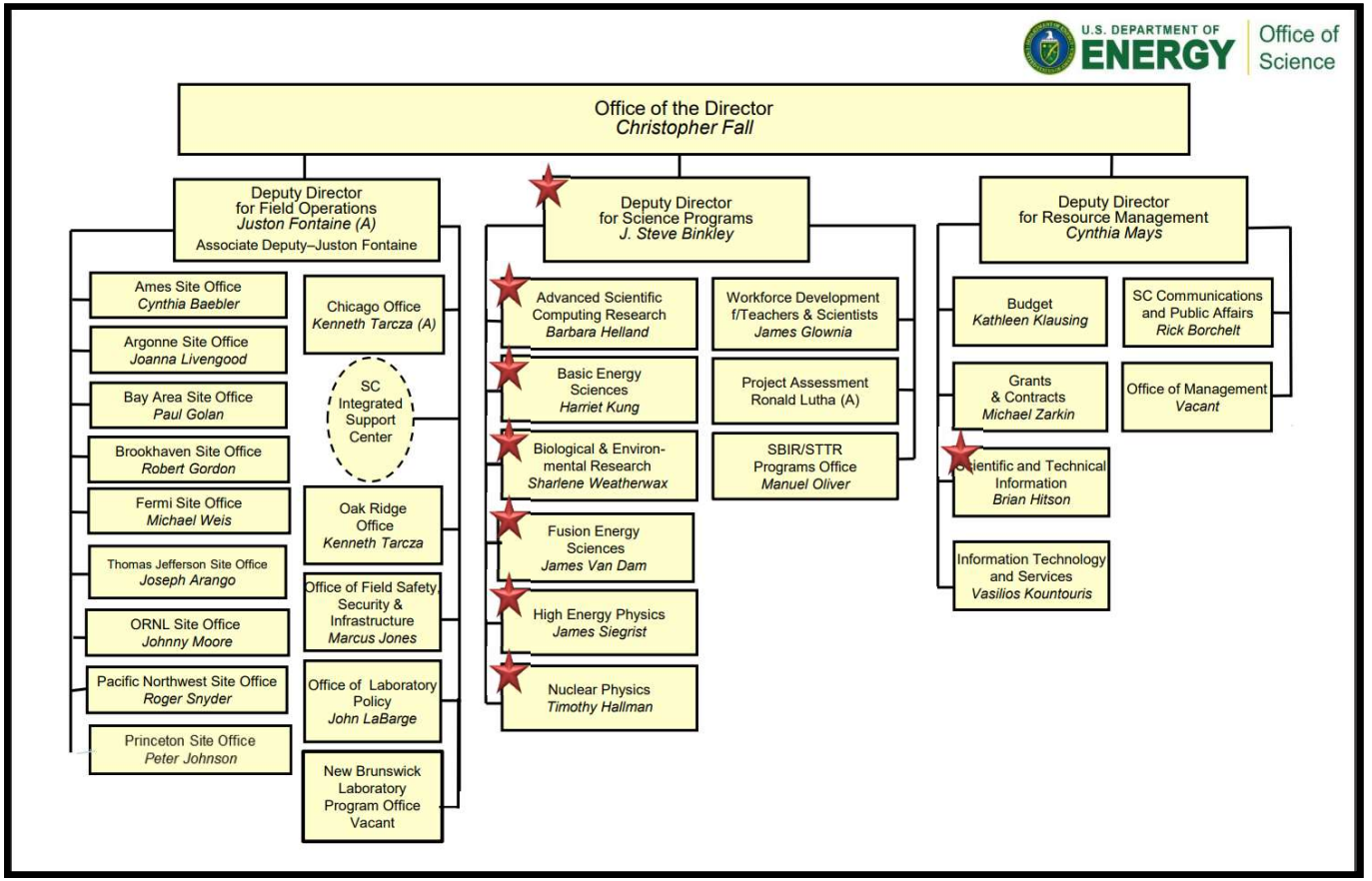
- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles
- I3. (meta)data include qualified references to other (meta)data

To be Reusable:

- R1. meta(data) are richly described with a plurality of accurate and relevant attributes
- R1.1. (meta)data are released with a clear and accessible data usage license
- R1.2. (meta)data are associated with detailed provenance
- R1.3. (meta)data meet domain-relevant community standards

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

Office of Science Working Group on Digital Data (SCWGDD)



★ SCWGDD Participants come from across SC