Federal Register Notice 86 FR 46278, https://www.federalregister.gov/documents/2021/08/18/2021-17737/request-for-information-rfi-on-an-implementation-plan-for-a-national-artificial-intelligence, October 1, 2021.

Request for Information (RFI) on an Implementation Plan for a National Artificial Intelligence Research Resource: Responses

Request for Information Response:
National Artificial Intelligence Research Resources Task Force

Abas Abdoli, Ryan N Coffee, Auralee Edelen, Michael Kagan, Daniel Ratner, Sohail Reddy, and Kazuhiro Terao
SLAC National Accelerator Laboratory

Applications of artificial intelligence and machine learning (AI/ML) to the sciences are already ubiquitous but have not reached their full potential, with scalability and access holding back researchers across the spectrum. This response focuses on **Question 2, Topic D**, concerning opportunities and specific steps that will enable AI/ML methods to increase their impact on the sciences and support researchers throughout the US.

A common thread among the most well-known applications of AI/ML in industry is the enormous scale of resources required. As an example, the Generative Pre-trained Transformer-3 (GPT-3) contains many billions of parameters, processed nearly a trillion words, cost more than $10M in compute, and required a dedicated team of ML and software engineers. Likewise in the sciences, one of the heralded AI/ML advances of the last year was the release of AlphaFold, again the product of a large, well-funded team from private industry with access to heavy resources in compute and engineering expertise and working with open scientific data. The scale of the AlphaFold model in combination with science domain knowledge proved a highly successful combination. While far smaller than GPT-3, AlphaFold is still larger than models commonly used in science today. While similar opportunities for AlphaFold-like models exist across the sciences, large-scale industry-built models are rarely directly applicable to scientific goals without significant modification and from-scratch retraining. It is vital to create ML/AI tools that are adapted to scientific data and tasks, and crucially are capable of scaling to the size where industrial AI/ML has seen break-throughs. Moreover, the real impact of AI/ML methods will appear when tools are available to all researchers, including students, individual contributors, and small academic groups across the breadth of research institutions in the US. Crucially, the tools must exist across the AI/ML lifecycle, including data collection, training, optimization, and deployment for automation and autonomous experiments. Widespread access to large AI/ML tools will both impact equity and innovation -- AI/ML tools should not be restricted to those lucky enough to have large resources -- and also will spread impact to scientists working on the most pressing scientific challenges of the day.

As scientists in the Department of Energy's national lab system, we see an opportunity to leverage existing investments in large-scale scientific experiments that can drive

AI/ML R&D broadly across the sciences. In particular, the DOE has invested in both computing resources and scientific facilities that generate the vast amounts of compute and data needed to train impactful AI/ML models. On the computing side, the current generation of leadership computing facilities at Argonne and Oak Ridge will push into the exascale while using AI/ML-friendly architectures. On the scientific-facility side, experiments are generating ever larger datasets. For example, the Vera Rubin Observatory will generate terabytes of data per day with its gigapixel camera, and SLAC's new x-ray laser facility will generate terabytes of data every second while operating around the clock. Even smaller scale instruments such as electron microscopes can produce 10 terabytes of images per day. The DOE's data and computing resources can be combined through the development of well-structured, benchmark datasets tied to specific science challenges, and paired with access to AI accelerating computational infrastructure like CPU clusters hosting farms of GPUs as well as emerging AI-tailored architectures such as developed by the likes of Google, Cerebras, Samba Nova, and GraphCore. Critically, access to datasets and such computing resources should be available to researchers across the spectrum of sciences and institutions.

Software Engineering: While computing and datasets are central to building AI/ML models, large-scale training also requires significant effort and expertise in software engineering. The requisite level of labor may not be available to individuals or small groups. We see several paths to solving this problem: creation of a software engineering workforce for AI/ML and science, investment in education, and development of open-access tools. At the national lab level, ML/software engineer teams could be leveraged across research groups to scale AI/ML projects. Democratizing engineering resources beyond national labs will be more challenging, but for example national engineering support for academic research could have a multiplicative effect on scientific progress. Any solution must be sufficiently flexible to support projects that vary from a small group of students to national collaborations with thousands of researchers. Investment in education is required both to create a dedicated workforce of software engineers as well as to educate scientists themselves. Researchers require training in both standard data science as well as in topics tailored to scientific AI/ML, such as uncertainty quantification, probabilistic models, and handling multi-modal data. Large-scale training -- requiring clusters and sophisticated data structures -- will be increasingly critical. Finally, open-access tools reduce the barrier to entry for novice AI/ML practitioners. Existing examples from industry such as PyTorch (from Facebook) and TensorFlow (from Google) have played an important role in the current democratization of AI/ML. Similarly, tools should be developed specifically for scientific challenges, for example handling new types of data, providing robust statistical analysis, or enabling large-scale training on clusters. Pre-trained models could also be

considered tools. For example, in the same way that the GPT-n family of pre-trained transformers was conceived as a generic tool for natural language processing, large-scale models pre-trained on domain-wide scientific datasets could be adapted to individual research tasks in new settings and with less labeled data in a sub-field with vastly reduced effort and investment.

Benchmark Datasets: The availability of benchmark datasets and associated models are key to democratizing AI/ML. However, shared infrastructure raises a host of issues to be addressed. In industrial applications, large-scale, public data (e.g. ImageNet, COCO, ShapeNet for Computer Vision) played a critical role in the breakthrough of modern AI/ML: benchmark datasets fuel the training of AI/ML models, enable transparent and fair comparison between solutions, and support interdisciplinary collaborations among domain experts. Access to benchmark datasets however remains a non-trivial challenge for most scientific datasets. Even within a single experimental collaboration, it may take years of training for new members to interpret data structures and develop tools necessary for a scalable AI/ML solution. Independent groups with no or limited access to the data and lacking knowledge of the data provenance are simply unable to participate. Furthermore, most tool development happens without a public or an inter-experimental collaboration in mind, thus requiring duplication of effort. An organized effort that encourages scientists to design databases and interfaces that enable efficient AI/ML development and sharing is urgently needed. Such a database should provide easy access (e.g. widely known data format, with community accepted visualization tools), detailed description of the contents (e.g. attribute descriptions, a complete metadata), definitions of scientific benchmark metric and test datasets, and direct coupling with data storage, distribution, and computing infrastructures behind the scene to allow efficient AI/ML model development. All of these components require support from software and data engineers as well as training resources for scientists to design effective solutions.

The goals enumerated here are challenging and the proposed solutions will require significant investment, but the potential for return on investment is enormous: new types of data analysis, improved and even autonomous operations and performance of instruments, and more. For the rest of this document, we give additional specific, actionable steps that relate to the topics in the RFI.

Question 1, Topic B, The appropriate agency responsible for the research resource:

It is recommended that the US Department of Energy (DOE) play a role in generating and hosting large-scale benchmark scientific datasets and associated AI models. Given

the number of scientific facilities operated by national laboratories paired with exascale computing resources, the DOE is in a prime position to contribute to the NAIRR effort. DOE facilities produce an overwhelming abundance of high-quality, multi-disciplinary datasets that span a wide range of scientific fields and span the spectrum from openly public to nationally secure datasets. These datasets include some of the largest scientific experiments operated by the federal government, as well as numerous smaller facilities in physics, chemistry, biology, materials, geology, environmental science, and more. On the computing side, the DOE is currently building the first two public exascale machines in the US -- Frontier at Oak Ridge National Laboratory and Aurora at Argonne National Laboratory -- both set to come online in 2022. In addition, the National Energy Research Scientific Computing Center (NERSC) is currently installing its own upgrade in Perlmutter. Predecessors (Summit and MIRA) may also be repurposed for NAIRR efforts, and additional computing systems, in particular IoT and Edge computing systems, exist throughout the DOE laboratory system. Due to the increasingly large size of scientific datasets -- commonly in the terabytes and reaching into the petabyte scale -- being generated in the DOE system, it is recommended that the computing resources be connected directly to the benchmark datasets and over network to the scientific instruments with the highest data production rates that are soon to exceed TB/second continuous operation.

Question 1, Topic C, Model for allocation of resources:

It is recommended that NAIRR initiative follow the tier-like allocation model currently used to allocate resources e.g. at ORNL on Summit. This model requires several computing architectures with variable levels of scalability and requires users to demonstrate the scalability/performance of their method/model on less intensive architecture before allocating high-tier resources. Such resource allocation may also be extended to ML engineering / software engineering support. Through an engineering workforce created through the NAIRR initiative and potentially in partnership with the private sector, users could request ML and software engineering support from a broad community. If a request meets the criteria set by the engineering review, it would be allocated engineering support. This engineering time could be requested on user grant applications. In addition to this model, it is recommended that the allocation of resources consider the following criteria:

1) Impact/significance of the research
2) Feasibility
3) Novelty
4) Benefits of research (in terms of public availability or inter-agency benefit of the resulting data/results, including equity)

5) Degree of collaboration across multiple institutions and/or federal agencies

Question 1, Topic D, Capabilities needed to create shared infrastructure

We have seen that both models and datasets can be cross-disciplinary, in the sense that a model or dataset from one domain can be adapted or applied in other domains. It is recommended that a set procedures/guidelines be developed that dictate the structure and interfaces for both models and datasets. For example, foundational architectures (e.g. GPT for text generation) can be tuned to perform a wide range of specific niche tasks within text generation, but more importantly for science these architectures can be re-trained from scratch and applied to completely new domains (e.g. music). To increase reusability and portability of previously developed models, a set of standards should be developed for model storage, training, deployment, etc. Similarly, a centralized set of standards for benchmark datasets in scientific domains is needed to govern data storage formats, access, and metadata to reduce engineering overhead and lower the barrier to training and comparing model performance. The chosen standards should follow the Findability, Accessibility, Interoperability, and Reusability (FAIR) data principles. As datasets are domain-specific, it is recommended that datasets are created and provided by domain experts and guided by standards and engineering tools to encourage FAIR principles. Such datasets can then be made accessible through NAIRR with appropriate access control. We expect that few datasets would need to be stored at the NAIRR facilities, but rather would be provided for remote access under a NAIRR licensing agreement.

Question 1, Topics E,F,G, Question 5: Limitation of NAIRR ability and Federated Machine Learning

A limitation of the NAIRR is the ability to generate and distribute data. Although data in science is abundant, at present only the private sector has the required data to develop AI for applied everyday use. The data gathering practices at Facebook, Google, Amazon, etc. allow the private sector to both develop and deploy their AI for mainstream uses. The involvement of NAIRR in gathering certain types of scientific data under a federal agency poses both privacy concerns and potentially national security concerns. The DoE is well positioned to apply security and encryption tools to data in a federal-scale AI initiative, and may actually be an opportunity to develop AI solutions that are more broadly applicable to e.g. financial services industry and healthcare.

Concerns regarding data and model dissemination (topic E) are fundamentally linked to both security (topic F) and privacy (topic G), and are as yet unaddressed by the scientific ML community. A Federated Machine Learning (FedML) ecosystem supports training across distributed datasets without need for direct sharing of data and would enable collaboration even among competing institutions. FedML leverages diverse datasets to ensure large models generalize rather than shoehorn into narrow niche applications. Large pre-trained models such as Bidirectional Encoder Representations from Transformers (BERT) and GPT-style transformers must be pre-trained on a broad distribution of training examples. For these language models, a tremendously diverse set of examples are required such as all of Wikipedia text. The resulting models can then be applied to niche problems with scarce labels. However, in domains where data is compartmentalized, for example due to privacy or national security concerns, it is not possible to assemble a single dataset of sufficient size to train a foundational architecture.

A well constructed FedML ecosystem could alleviate this issue with e.g. homomorphic encryption based ML training whereby loss computation neither exposes the model weights nor data samples of the host repository. This could allow for even niche science cases to contribute to and benefit from a generalizable base model. Already, the financial services industry has begun to derive market value from FedML architectures in which competing institutions leverage shared knowledge while still preserving their individual competitive advantage. We feel that such a model for science would allow for inter-domain and even inter-lab competition while nevertheless encouraging an integrated web of robust and evolutionarily improving scientific base models without ever exposing sensitive or private data.