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

DISCLAIMER: Please note that the RFI public responses received and posted do not represent the views and/or opinions of the U.S. Government nor those of the National AI Research Resource Task Force., and/or any other Federal agencies and/or government entities. We bear no responsibility for the accuracy, legality, or content of all external links included in this document.

BERKELEY LAB

October 1st, 2021

Via email: NAIRR-responses@nitrd.gov

White House Office of Science and Technology Policy and National Science Foundation

Wendy Wigen, NCO,

2415 Eisenhower Avenue,

Alexandria, VA 22314

RFI Response: National AI Research Resource

Thank you for the opportunity to respond to the National AI Research Resource (NAIRR) RFI. As one of DOE's national laboratories, Berkeley Lab specializes in integrative science and technology, taking advantage of our world-renowned expertise in materials, chemistry, physics, biology, earth and environmental science, mathematics, and computing. We advance the frontiers of science and technology through three approaches: advanced instrumentation and user facilities, large team science, and core research programs led by outstanding investigators.

In 2016 we launched an Artificial Intelligence for Science initiative, which quickly developed into one of our most far-reaching and successful initiatives ever. The initiative took advantage of our strength in applied mathematics and computer science and combined that with our expertise in basic science across all scientific disciplines. The website http://ml4sci.lbl.gov describes some of the resulting applications in which AI techniques have been successfully applied to cosmology, particle and nuclear physics, materials science and engineering, chemistry, synthetic biology, genome and biomedical sciences, environmental biology, geoscience, watershed science, climate modeling, technology scale-up, smart grid, water treatment, transportation, advanced detectors, accelerator operations, and many more. Based on our experiences developing and using AI methods and our expertise in managing scientific data through the entire lifecycle, we have the following recommendations.

A National Artificial Intelligence Research Resource must use a holistic approach to combine elements for computation, data lifecycle management, user interface, and training. (Response to Question 1-A)

The stated goal for the NAIRR is to democratize access to the cyberinfrastructure that fuels AI research and development, enabling all of America's diverse AI researchers to fully participate in exploring innovative ideas for advancing AI, including communities, institutions, and regions that have been traditionally underserved—especially with regard to AI research and related education opportunities. To achieve this goal, the NAIRR must include much more than a set of loosely coordinated computational and data storage resources, but also devote ample attention to providing the tools, technologies, networking, and training to enable researchers to develop and deploy AI methods. Going beyond lip service to the "data lifecycle" to enable truly full-scope lifecycle development from acquisition to preparation to interface is essential. At Berkeley Lab, we have found that multidisciplinary teams composed of experts in AI, user interface design, distributed systems, security, and hardware, in addition to experts in the scientific domain, have been essential to building systems and tools that can push the boundaries of AI research.

- The most impactful AI solutions often require large-scale computing which remains inaccessible for many researchers. New methodologies need to be developed to allow individual researchers to *easily* exploit distributed and heterogeneous compute resources for AI model development. Doing so would democratize large-scale AI, enabling effective use by a broad research community.
- Advanced networking is crucial for democratization of access to key data sets across institutions large and small, urban and rural, and domestic and international.
- There are many barriers preventing researchers from accessing conventional HPC or commercial cloud resources, for example, lack of technical know-how on how to scale from laptop resources to extreme concurrencies, or how to package a workflow into a container, etc., that are not solved simply by providing resources without technical help that will provide on-ramps for researchers with varying levels of skill. The NAIRR needs a plan to provide this.
- There is a need for tools to evaluate the appropriateness and limitations of AI models when applied to research questions, as well as automated model selection and architecture design. Automated model selection will also help address democratization for researchers less well versed in AI. Researchers should also have a credible path to reach out to AI specialists to get their problems addressed.
- It is crucial that the NAIRR allow for models of AI deployment where the workflow is tightly coupled. A key example here is the case of automated experiments. For example, self-driving labs are being proposed for synthetic biology, the operation of beamlines, materials synthesis, earth observations and many others. For synthetic biology, this approach leverages microfluidic chips coupled with real-time DRL models to automate cell DNA modification to improve biofuel yields as part of the "Design Build Test Learn" (DBTL) cycle. The automation can substantially increase exploration of the combinatorially complex state space over current, largely-manual methods, and democratize science by automating a process where human scientific labor is less available.
- The NAIRR should partner with the broader community on leveraging existing and developing additional open standards for AI software and data, and implement these standards for tools developed as part of the NAIRR. This will help ensure researchers can migrate from using the NAIRR to other research-focused resources, or commercial

cloud providers. In addition, standards could include AI benchmarks for evaluating performance on current and emerging architectures.

- As the NAIRR evolves there should be a mechanism for soliciting and incorporating user feedback so that its design and implementation can incorporate diverse perspectives and emerging trends in the field. The NAIRR could also act as a vehicle for user-experience research to improve the overall effectiveness of the resource.

AI software will need to deeply integrate with simulations, data analysis pipelines and legacy codes (Response to Question 1-D and 2-D)

Unlocking the transformative potential of AI in research will require deep coupling of AI and traditional simulation or data analysis applications; as well as in steering and tuning of experimental/simulation workflows with AI. To enable this the NAIRR should recognize:

- R&D will be required to determine both performant and scalable methods for coupling of AI with experiment and simulation, as well as the mix of computational resources to effectively run this mixed workload, including incorporating potential novel AI hardware.
- Research and science communities often have large legacy code bases, written in a mix of programming languages, as well as complex workflows and ad-hoc curation of datasets. Investments should be made to support integration of these with current AI software, such as the python-based deep-learning frameworks.

Data will need to be made *AI-ready* (Response to Question 1-D and 2-D)

The quantity and *quality* of data available to build AI models is a driving determiner of the resulting model's performance. These datasets are dynamic, continually being improved and extended as new data and metadata become available, all the while incorporating expert knowledge. Furthermore, the development of AI models needs to be able to inform the data collection and curation process to enable improvements.

To address these requirements the NAIRR should:

- Federate data centers and partner with the experts that maintain the needed datasets in a shared research infrastructure to enable the AI models to be built using the latest data and expertise.
- Ensure that the needs of AI are incorporated into initial data collection as well as throughout the data lifecycle, including avoiding bias.
- Develop systems that enable tracking of the data and version used in the development of AI systems to allow for retraining and understanding of the limitations of models built from the data.
- Ensure AI model development is part of a cycle of continuous improvement that includes the data collection, processing, and curation needed to drive the models. It will be critical that there is feedback to the data collection and generation process to gather the necessary data and metadata for AI.

- Develop benchmark datasets and challenges that specifically target, and push development of, AI features required for research and science. For example, those that allow for quantifying uncertainty and for the coupling of AI with simulation and data analysis.

Data owners/providers should be able to make data available for AI-based research without requiring full trust of data and computing centers and scientific end-users. (Response to Questions 1-D, 1-F, 2-D, and 2-F)

Two drivers of data owners' reluctance to share data are the risks of sharing sensitive data, and the risks of hosting such data. Even with strong security protections, traditional enclaves still require implicitly full trust in the facility hosting the sensitive data, thereby increasing the liability of an institution for accepting responsibility for hosting data. This limitation can significantly weaken the trust relationships involved in sharing data, particularly when groups are large and distributed. Also, traditional security protections often hinder analysis processes for the scientific community whose abilities and tools are optimized for working in open, collaborative, and distributed environments.

Emerging hardware security technologies, including hardware trusted execution environments (TEEs) can form the basis for platforms that provide strong security benefits while maintaining computational performance, without requiring that system administrators at computing and data centers be fully trusted. Commercial TEEs from the major CPU vendors exist, and have been adopted by the major commercial cloud vendors. To address security concerns of sensitive data, the NAIRR should:

- Leverage TEEs to provide strong security isolation guarantees to protect sensitive data, even from malicious system administrators.
- Support research and development to enable future TEEs to continue to improve both performance and security over today's commercial TEEs to enable a broader range of secure scientific AI applications.

Differential privacy is a statistical technique that can put bounds on the amount of information about a dataset that can be leaked to a data analyst as a result of a query or computation by adding "noise" and enforcing a "privacy budget." It has emerged as an approach to provide strong privacy protection of data output and is now a mainstream solution, with production use by Apple, Google, and the U.S. Census Bureau, the existence of several open source distributions, and successful application to a diverse range of data types. To address privacy and confidentiality concerns of sensitive data, the NAIRR should:

- Leverage differential privacy techniques to enable analysis and AI model training while limiting private information leakage.
- Support research and development to advance the usability of differential privacy and integration of differential privacy in scientific workflows so it can more easily be broadly leveraged for large-scale, scientific AI applications.

Existing Berkeley Lab Resources (Response to Question 4)

- The Department of Energy Office of Science High Performance Computing User Facility NERSC, houses computing and storage resources, including the >120Petaflop Perlmutter supercomputer, with over 6000 NVidia A100 GPUs, a ~120PB Community File System and a 225PB Archival storage system, and data portals for over 8000 Office of Science researchers. NERSC also evaluates next generation AI hardware and testbeds, as well as playing a lead role in developing benchmarks tailored for the research community and for HPC-scale with the MLPerf HPC working group.
- ESnet provides the high-bandwidth, reliable connections that link scientists at national laboratories, universities, and other research institutions, enabling them to collaborate on some of the world's most important scientific challenges including energy, climate science, and the origins of the universe. Funded by the DOE Office of Science, ESnet is managed and operated by the Scientific Networking Division at Lawrence Berkeley National Laboratory. As a nationwide infrastructure and DOE User Facility, ESnet provides scientists with access to unique DOE research facilities and computing resources.
- Berkeley Lab provides expertise and software enabling automated cross DOE SC facility research and collaborations including data analysis from light sources, telescopes, sensors, sequencers and microscopes.
- The lab has demonstrated AI research applications running at increasingly large computing scales, and with increasingly sophisticated approaches (including the 2018 Gordon Bell Prize winner).
- The user facilities and data science platforms at Berkeley Lab - the Joint Genome Institute, the Department of Energy Systems Biology Knowledgebase (KBase), the Advanced Light Source, the Materials Project, and the Molecular Foundry - are premier data generators that fuel new scientific machine learning solutions. In particular, the Materials Project accelerates materials discovery through its comprehensive database of materials properties generated through advanced simulations. The current data set includes >150,000 entries and is used by 150,000 registered users worldwide. The Materials Project has been used to develop many new ML methods, both internally and by the larger research community, for predicting the properties of new chemical compounds.
- NERSC and Berkeley Lab host a number of data sets drawn from experiments in high-energy physics and cosmology, including imaging and spectroscopy for large cosmological surveys seeking to map the effects of dark energy (BOSS, eBOSS, DESI, SN Factory, PTF and ZTF), dark matter search data (LUX, LZ), reactor neutrino data collected over ten years by the Daya Bay experiment, and cosmic microwave background data from more than a dozen ground-based and satellite experiments. We also host the Particle Data Group, a DOE Office of Science Public Reusable Research (PuRe) Data resource, which provides an authoritative, curated database comprising all published

results in particle physics and aspects of cosmology, as well as world averages, uncertainties, and reviews that explain the underlying physics.

- NERSC and Berkeley lab host the ESS-DIVE data repository that contains diverse environmental datasets spanning observational, experimental and modeling research. ESS-DIVE allows data contributors to archive, manage and share various types of data in standardized formats, and obtain digital object identifiers that can be used to cite and track usage of the data. ESS-DIVE users are able to find and obtain data generated by ESS researchers that is organized for better interpretation, analysis, and integration.
- Berkeley Lab runs educational events focussed on the science and research community, including the Deep Learning for Science Summer School and Webinar Series, and practical training for running at HPC scale such as tutorials run by NERSC in collaboration with Intel, Cray/HPE and Nvidia at the Supercomputing conference since 2018.
- Berkeley Lab, with the proximity to University of California and the broader Bay Area has the ability to train students and interns at scale through well-defined internship programs and community outreach. This includes partnerships with the Sustainable Horizons Institute that explicitly seeks to increase representation of students and researchers from marginalized or disadvantaged backgrounds in science.

Public-private partnerships will be the fastest route to resilience and availability for the NAIRR (Response to Question 5)

Over the last 10 years, commercial cloud-computing and storage resources have grown to rival or exceed federally funded high performance computing resources in sheer core count. While HPC resources still often have advantages such as better interconnects and I/O subsystems, the commercial cloud offers resiliency, accessibility, and the ability to surge from using small to larger amounts of resources. As a way to build up capability quickly, a national resource could be composed as an array of compute and data resources, composed of existing high-performance computing resources, or extensions to resources at existing HPC Centers, commercial cloud options, and a layer of software tools. As the resource matures, continued use of commercial cloud resources in addition to dedicated HPC resources makes sense to provide flexibility to cope with surges in demand, and the extensive geographical distribution of cloud resources will additionally provide for more uniform access to data and compute nationally.

In addition, many of the commercial cloud providers have been extremely active in developing artificial intelligence frameworks and other software tools. Often, these frameworks are open-source, and are used by a broad audience of scientists. Here, the NAIRR will need to take care to avoid researchers being locked into proprietary frameworks and datasets that cannot be migrated to other AI resources.

Similarly, multiple datasets have been curated and made available to the wider community by commercial entities with analytical capabilities, e.g. Google Earth Engine, which have been a boon to researchers lacking the resources to do this by themselves. Expanding such resources as part of the NAIRR is another beneficial part of a public-private partnership.

The COVID-19 HPC Consortium could provide a model for how a public-private partnership might function. The Consortium was spun up quickly in response to the demand for computational modeling in the face of the COVID-19 panic, and consisted of cloud and HPC platforms provided by IBM, Microsoft, IBM, NSF HPC Centers, and the DOE National Laboratories. Representatives from each institution formed a governing body to set basic policy, and separate committees were set up to evaluate proposals submitted by researchers applying for access. Successful proposals were directed to the most appropriate resource.

Respectfully submitted,

Deb Agarwal, Division Director, Scientific Data Division
Katie Antypas, Division Deputy, National Energy Research Scientific Computing Division
Wahid Bhimji, Big Data Architect, National Energy Research Scientific Computing Division
Jonathan Carter, Associate Laboratory Director, Computing Sciences Area
Sean Peisert, Staff Scientist, Scientific Data Division
Charuleka Varadharajan, Research Scientist, Earth and Environmental Sciences Area
Peter Zwart, Staff Scientist, Molecular Biophysics and Integrated Bioimaging Division

Contact:
Jonathan Carter
[redacted]