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.

OSTP and NSF Implementation Plan for a National Artificial Intelligence Research Resource (NAIRR)

Northeastern University Response

Organization Information:

Name/Address of Organization: Northeastern University, 360 Huntington Ave, Boston, MA 02120
Business Size of Organization: 6,307 employees (as of Fall 2020)
Socio-economic status of Organization: 501(c)3
Northeastern Respondents: Jennifer Dy, Deniz Erdogmus, Usama Fayyad, David Kaeli, Ningfang Mi, Dana DeBari, Timothy Leshan

Northeastern University Responses:

1) What options should the Task Force consider for any of roadmap elements A through I above, and why? [Please take care to annotate your responses to this question by indicating the letter(s) of the item (A through I in the list below) for which you are identifying options.]

The performance of AI and machine learning (ML) algorithms are highly dependent on the quality and size of available data. Training AI/ML models for large data requires higher computational and storage resources. Moreover, modern powerful ML models (deep learning, large language models, high-dimensional and sparse data sets) are increasingly more complex, requiring an increasing number of learnable parameters and repeated iterations – often requiring "bursty usage" of elastic computing grids.

The availability of massive data and computational resources is crucial in making progress in AI/ML, which is mostly available only to large private companies and national labs. Component D is one of the most critical components to have for a National AI Research Resource (NAIRR). Another important component is H to continue to support research in AI through federal grants in partnership with the private sector and national labs. See responses below for details.

Component D is also potentially very relevant to addressing two of the biggest bottlenecks facing any organization that wants access to cloud elastic computing:

a. The difficulty of setting up a cloud environment (see further discussion below)
b. The expense for data egress costs once a public or private cloud environment is up and running – while the cost of storage is cheap, the cost of moving clouds or moving data to other computational environments becomes prohibitive. Creating a shared storage cluster by NAIRR that allows high-speed low-cost transfers of data out of clouds would address a big issue and would create more competition between the big cloud providers.

Component F on data security is gaining prominence, especially when sharing data on a limited basis. Many organizations in finance, healthcare, and telecom are learning that the best way to "share data" without data leakage to potentially malicious organizations and nation states is by creating environments where the participants bring their applications to the data rather than bring data to the applications in the standard typical model. A compute cloud where data is stored centrally, and users are enabled to run their algorithms on the data without moving it (via data lakes or storage clusters coupled with significant compute) creates a more tenable approach to controlling data movement and unauthorized distribution of data. NAIRR can play a central role to enable such a model to ensure that data is only used by authorized parties and no indirect paths to share "copies" are allowed. This also creates strong opportunities to impose privacy-preserving mechanisms (e.g., access via differential privacy) to further guarantee responsible use (Items G and F).

2) Which capabilities and services (see, for example, item D above) provided through the NAIRR should be prioritized?

Over the past half-century, the United States has made significant investments in high-performance computing (HPC). These centers have designed powerful computing infrastructures that serve the computing needs of thousands of scientists and engineers as they generate new discoveries and remedies, enabling us to develop a better understanding of our world.

In 2010, high-performance computing (HPC) researchers from Lawrence Berkeley National Laboratory observed that cloud computing and data analytics were starting to dominate the data center, and so should be evaluated as a potential platform for HPC [Guida2020]. The study found that many scientific applications could be effectively scaled on cloud-based HPC platforms, but major impediments remained to fully leverage the Cloud for HPC applications, including memory system performance and communication costs. In 2015, Reed and Dongarra argued that the future of scientific discovery would combine advances from both machine learning and traditional HPC, producing a new and novel high-performance computing ecosystem [Reed2015]. Due to major differences in the characteristics of these workloads, it becomes ever more challenging to deliver a single computing framework. Today, many machine learning workloads are designed to leverage containerized services [Wang2018, Jouppi2020, Yi2020] and cloud-based hardware/software stacks [Cusumano2019] to process the growing volumes of data.

Elastic Cloud technology can provide an effective model that can deal with the heterogeneity issues associated with emerging HPC and machine learning applications. One major motivation for considering co-hosting these two classes of workloads is that many HPC applications generate mountains of data that need to be explored [Buffat2014, Balaprakash2019]. Further, HPC applications are beginning to leverage machine learning algorithms to efficiently steer simulations and iterative applications, leading to solutions in a fraction of the time [Wozniak2018, Kurth2018, Dong2020]. We believe that this emerging model will play an important part as we develop a national infrastructure to support exploration in artificial intelligence. This can be thought of as the next generation adaptive computation for delivering optimized AI workloads platforms – reducing time, cutting energy usage for sustainability, and maximizing throughput across more workloads.

There is a growing number of machine learning applications that require large-scale data access and processing. Scale-out storage infrastructure has become essential for storage systems (i.e., hyperscaler [Lu2018] and cloud-storage [Luo2020]) to provide vast space and high throughput. One promising direction of solutions is to deploy a disaggregated model on cloud-based ML platforms where storage drives are physically separate from the compute nodes. In such a system, compute and storage resources can be scaled independently for different needs, and resource management becomes more flexible. All-flash arrays have emerged in data centers recently, yielding superior performance to traditional storage devices. A hybrid of diverse storage devices, including traditional HDDs, SSDs with PCIe interface (e.g., NVMe), and phase-change memory devices (e.g., Intel Optane), can be installed in storage nodes, providing a variety of device capacities and processing speeds. Various scheduling schemes and management policies can further be applied on a single or a group of compute/storage nodes based on the demands of machine learning workloads. Another possible motivation for disaggregating the storage system is that the safety and integrity of (sensitive) data can be maintained by applying different data preservation policies on different storage nodes, which thus uncouples the access to private and public data repositories.

3) How can the NAIRR and its components reinforce principles of ethical and responsible research and development of AI, such as those concerning issues of racial and gender equity, fairness, bias, civil rights, transparency, and accountability?

As AI is widely deployed in a variety of applications affecting various sectors of society, we need to make sure AI systems are trustworthy: i.e., ethical, responsible, fair, reliable, secure, privacy preserving, safe, transparent, and interpretable. Research in trustworthy AI is in its infancy requiring convergence of experts from multiple disciplines, such as philosophy, law, sociology, psychology, and AI. To help advance this new field in AI, NAIRR through component H supporting federal funding in partnerships with the private sector can help accelerate development of this field. NAIRR can also help by establishing an appropriate agency that provides guidelines for ethical oversight. In designing a shared data and computing resource, NAIRR can help reinforce principles of ethical and responsible AI through components F (on security) and G (on privacy).

Furthermore, the approach to provide equitable access to computation and analytics is to work on lowering the bar for skills needed to access the technology. Leveraging NAIRR-class resources requires a high degree of technical skills, specialized talent and know-how, and the use is far from user-friendly. We need to reduce these barriers of entry and make these resources accessible to much larger groups of users by simplifying their use and by embedding pre-prepared packaged “solutions” to many of the common problems in analytics and Data Science. Usage can be increased by also making the outputs (e.g., analytics, insights, reports, etc.) much easier to produce and interpret. This simplification is much needed not only to increase the user base and make access more equitable, but to also reduce the significant redundant effort by advanced groups in building up the tools, utilities and environments to make their own work easier.

4) What building blocks already exist for the NAIRR, in terms of government, academic, or private-sector activities, resources, and services?

Two unique resources that are available at Northeastern are the DARPA Colosseum Network Simulator and the membership in the Massachusetts Green High Performance Computer Center (MGHPCC). Each of these facilities provide unique capabilities to the AI/ML research community.

NAIRR will need to acquire and provide AI-ready data sets. This is a new requirement that is needed by the AI/ML research community. Rich data sets from key domains (e.g., health, the exposome, 5G/6G communications, severe weather, the environment, etc.) will attract researchers from a broader range of areas to leverage the capabilities of NAIRR. We will need appropriate models for facilitating data use agreements, ensuring data security and privacy, while also considering the fairness and bias issues related to these data sets. With the major investment by Northeastern in the Institute for Experiential AI (focused on working with partners in its AI Solutions Factory and on Responsible AI practice), and other institutes in Robotics, Cybersecurity, Network Science, and Wireless IoT, Northeastern has many galvanized research and application resources dedicated to solving the critical problems to society.

In practice, much research work in ML that enables AI is evaluated and published over unrealistically simple data sets and schema. Real world data is much more complex: data schema and the variations in data quality and availability. Most such data sets are not available to the general research community. Thus, much of their work is of little relevance to real applications. However, there is no lack of data. The issue is the expense of gathering it, cataloging it, labelling it, etc. There is a tremendous amount of data available on the Web. However, crawling it, performing entity extraction, and then organizing and labelling it properly is prohibitively expensive. Delivering this service through NAIRR will help to generate the next generation of challenge data sets and will go a long way in advancing research and in removing the obstacles that are perceived as insurmountable by researchers in Machine Learning and Data Science. Many such training sets can be also from simulations and outputs of the HPC applications, in addition to traditional web and social media sources. Many other complex and large publicly available data sets can be leveraged: financial markets, capital and equity markets, census data, environmental data, economic data, etc. The problem facing most researchers is getting this data ready for analysis, which is a huge challenge.

5) What role should public-private partnerships play in the NAIRR? What exemplars could be used as a model?

At the Massachusetts Green High Performance Computer Center, we have aggressively explored public/private partnerships in both the initial buildout of our Center, and as a theme for ongoing research into public/private cloud services. Specifically, the Center was founded as a partnership between the State of Massachusetts and five Massachusetts universities (Boston University, Harvard, MIT, Northeastern and the University of Massachusetts system). Each partner contributed to build a state-of-the-art green computing facility that services researchers across the Commonwealth.

In addition to the creation of this facility, ongoing research at the MGHPCC has aggressively pursued unique public-private opportunities that leverage the Center. Two examples of this focus include the Massachusetts Open Cloud (MOC), an ongoing project that has created a self-sustaining at-scale public cloud based on the Open Cloud eXchange model [MOC]. The MOC serves as a marketplace for industry partners, as well as a service for researchers and industry to innovate and expose innovation to real users. The second example of a successful public-private engagement is the AI Jumpstart program. This state-funded project leverages the expertise in AI/ML of three leading academic institutions (Northeastern, Boston University, and Tufts) with small- and medium-sized businesses that are looking to leverage the benefits of AI/ML technology within their organizations [AI-Jumpstart]. The support from the state provides for acquisition of a large computing cluster specifically designed for the most challenging AI/ML workloads. The state funding supports initial engagement grants for the companies to use to work with the faculty and their students.

Scaling programs such as AI Jumpstart to the national level and creating a framework where many companies are extremely interested in innovating with AI and data, though are not able to because of lack of platforms and know-how, would be a huge contributor to the economy. Such a framework would enable academia and small and medium-sized companies, who have the talent and skills, to work with many companies and organizations who do not. This is akin to the building of railroads in the 1800's and the interstate highway system in the 1900's – both enabled tremendous economic expansion and opportunities – we need a similar infrastructure around data, AI and computation.

6) Where do you see limitations in the ability of the NAIRR to democratize access to AI R&D? And how could these limitations be overcome?

For NAIRR to be successful in terms of meeting the needs of a broad spectrum of researchers and scientists, two major hurdles need to be addressed. The first is the cost of services, which the NAIRR framework should help to address in part. How do we continue to provide access to this infrastructure to everyone, and ensure that the facility can meet the needs of those with fewer financial resources? One model is to charge a fee to those institutions that have financial resources and utilize this fee to ensure there are ample resources reserved for those institutions that have limited funding. A fair and equitable financial model to accompany this facility would have to be developed and deployed, one that ensures its long-term vitality and sustainability.

A second barrier that needs to be overcome is to ensure that everyone receives a proper education on how to best use this infrastructure. This includes the effective use of middlewares and tools/libraries that support effective use of these resources by non-specialists. In terms of hosting this infrastructure, expenditures on the equipment and services should come with an equal amount of support to deliver training to the future users of this infrastructure. Academic institutions with strong programs in AI and Data Science seem well-poised to deliver these services.

References:

[Guida2020] Giulia Guidi, Marquita Ellis, Aydin Buluc, Katherine Yelick, and David Culler. 10 Years Later: Cloud Computing is Closing the Performance Gap, 2020.

[Reed2015] Daniel A. Reed and Jack Dongarra. Exascale Computing and Big Data. Communications of the ACM, 58(7), 2015.

[Wang2018] Naigang Wang, Jungwook Choi, Daniel Brand, Chia-Yu Chen, and Kailash Gopalakrishnan. Training Deep Neural Networks with 8-Bit Floating Point Numbers. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18, page 7686–7695, Red Hook, NY, USA, 2018. Curran Associates Inc.

[Jouppi 2018] Jouppi, Norman P. and Yoon, Doe Hyun and Kurian, George and Li, Sheng and Patil, Nishant and Laudon, James and Young, Cliff and Patterson, David. A Domain-Specific Supercomputer for Training Deep Neural Networks. Commun. ACM, 63(7):67–78, June 2020.

[Yi2020] Yi, Xiaodong and Luo, Ziyue and Meng, Chen and Wang, Mengdi and Long, Guoping and Wu, Chuan and Yang, Jun and Lin, Wei. Fast Training of Deep Learning Models over Multiple GPUs. In Proceedings of the 21st International Middleware Conference, Middleware '20, page 105–118, New York, NY, USA, 2020. ACM.

[Cusumano2019] Michael A. Cusumano. The Cloud as an Innovation Platform for Software Development. Communications of the ACM, 62(10):20–22, September 2019.

[Buffat2014] Marc Buffat, Lionel Le Penven, and Anne Cadiou. High Performance computing and Big Data for turbulent transition analysis. http://www.netlib.org/utk/people/JackDongarra/CCDSC-2014/talk15.pdf, 2014. Online.

[Balaprakash2019] Prasanna Balaprakash, Romain Egele, Misha Salim, Stefan Wild, Venkatram Vishwanath, Fangfang Xia, Tom Brettin, and Rick Stevens. Scalable reinforcement-learning-based neural architecture search for cancer deep learning research. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, pages 1–33, 2019.

[Wozniak2018] Justin M. Wozniak, Rajeev Jain, Prasanna Balaprakash, Jonathan Ozik, Nicholson T. Collier, John Bauer, Fangfang Xia, Thomas S. Brettin, Rick Stevens, Jamaludin Mohd-Yusof, Cristina Garcia- Cardona, Brian Van Essen, and Matthew Baughman. Candle/supervisor: a workflow framework for machine learning applied to cancer research. BMC bioinformatics, 19(18):491, 2018.

[Kurth2018] Thorsten Kurth, Sean Treichler, Joshua Romero, Mayur Mudigonda, Nathan Luehr, Everett Phillips, Ankur Mahesh, Michael Matheson, Jack Deslippe, Massimiliano Fatica, et al. Exascale deep learning for climate analytics. In SC18: International Conference for High Performance Computing, Networking, Storage and Analysis, pages 649–660. IEEE, 2018.

[Dong2020] Wenqian Dong, Zhen Xie, Gokcen Kestor, and Dong Li. Smart-pgsim: using neural network to accelerate AC-OPF power grid simulation. arXiv preprint arXiv:2008.11827, 2020.

[MOC] The Massachusetts Open Cloud, URL: https://massopen.cloud/

[AI-Jumpstart] Providing Massachusetts Businesses with an AI Jumpstart, URL:https://innovation.masstech.org/AIJumpstart

[Lu2018] X. Lu, J. Chiu, S.-J. Chao, and Y.-B. Ye, "Design of instruction analyzer with semantic-based loop unrolling mechanism in the hyperscalar architecture," in ICS, 2018.

[Luo2020] S. Luo, G. Zhang, C. Wu, S. Khan, and K. Li, "Boafft: Distributed deduplication for big data storage in the cloud", IEEE Transactions on Cloud Computing, vol. 8, pp. 1199–1211, 2020.