LESSONS LEARNED FROM FEDERAL USE OF CLOUD COMPUTING TO SUPPORT ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT

Report by the
MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE SUBCOMMITTEE

of the
NATIONAL SCIENCE AND TECHNOLOGY COUNCIL

July 2022

About the Office of Science and Technology Policy

The Office of Science and Technology Policy (OSTP) was established by the National Science and Technology Policy, Organization, and Priorities Act of 1976 to provide the President and others within the Executive Office of the President with advice on the scientific, engineering, and technological aspects of the economy, national security, homeland security, health, foreign relations, the environment, and the technological recovery and use of resources, among other topics. OSTP leads interagency science and technology policy coordination efforts, assists the Office of Management and Budget with an annual review and analysis of Federal research and development (R&D) in budgets, and serves as a source of scientific and technological analysis and judgment for the President with respect to major policies, plans, and programs of the Federal Government. More information is available at https://www.whitehouse.gov/ostp.

About the National Science and Technology Council

The National Science and Technology Council (NSTC) is the principal means by which the Executive Branch coordinates science and technology policy across the diverse entities that make up the Federal R&D enterprise. A primary objective of the NSTC is to ensure science and technology policy decisions and programs are consistent with the President's stated goals. The NSTC prepares R&D strategies that are coordinated across Federal agencies aimed at accomplishing multiple national goals. The work of the NSTC is organized under committees that oversee subcommittees and working groups focused on different aspects of science and technology. More information is available at https://www.whitehouse.gov/ostp/nstc.

About the Machine Learning and Artificial Intelligence Subcommittee

The Machine Learning and Artificial Intelligence (MLAI) Subcommittee monitors the state of the art in machine learning (ML) and artificial intelligence (AI) within the Federal Government, in the private sector, and internationally to watch for the arrival of important technology milestones in the development of AI, to coordinate the use of and foster the sharing of knowledge and best practices about ML and AI by the Federal Government, and to consult in the development of Federal MLAI R&D priorities. The MLAI Subcommittee reports to the NSTC Committee on Technology and the Select Committee on AI.

About This Document

This document aims to capture lessons learned from the activities spearheaded by various agencies to enhance access to cloud computing resources to advance federally funded AI R&D and highlight potential opportunities going forward for optimizing Federal use of commercial cloud as a component of broader efforts to further AI R&D that can accelerate scientific discovery and address societal challenges. This report focuses specifically on the progress Federal departments and agencies are making pursuant to the directive in EO 13859 to prioritize allocation of high-performance computing resources for AI.

Acknowledgments

The MLAI Subcommittee gratefully acknowledges the work of the Networking and Information Technology R&D (NITRD) Subcommittee, NITRD IWGs, NITRD National Coordination Office staff, and Future Advanced Computing Ecosystem (FACE) Subcommittee for their contributions to this report.

Copyright Information

NATIONAL SCIENCE AND TECHNOLOGY COUNCIL

Chair

Alondra Nelson
Deputy Director for Science and Society, Performing the Duties of Director, Office of Science and Technology Policy

Acting Executive Director

Kei Koizumi
Principal Deputy Director for Policy, Office of Science and Technology Policy

MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE SUBCOMMITTEE

Co-Chairs

Stephen Binkley, Principal Deputy Director, Office of Science, Department of Energy

Erwin Gianchandani, Assistant Director for Technology, Innovation and Partnerships, National Science Foundation

Lynne Parker, Deputy Chief Technology Officer of the United States

Elham Tabassi, Chief of Staff, Information Technology Laboratory, National Institute of Standards and Technology

Executive Secretary

Faisal D'Souza, Networking and Information Technology R&D, National Coordination Office

Writing Team

Gil Alterovitz, Department of Veterans Affairs

Rupak Biswas, National Aeronautics and Space Administration

Patricia Flatley Brennan, National Institutes of Health

Tess DeBlanc-Knowles, Office of Science and Technology Policy

Erwin Gianchandani, National Science Foundation

Nikunj Oza, National Aeronautics and Space Administration

Lynne Parker, Office of Science and Technology Policy

Venkatachalam Ramaswamy, National Oceanic and Atmospheric Administration

Dominic Sale, General Services Administration

Eddie Tejeda, General Services Administration

Table of Contents

Abbreviations and Acronyms

AI	Artificial Intelligence
API	Application Programming Interface
EO	Executive Order
FACE	Future Advanced Computing Ecosystem
GPU	Graphics Processing Unit
IWG	Interagency Working Group
ML	Machine Learning
MLAI	Machine Learning and Artificial Intelligence (Subcommittee)
NASA	National Aeronautics and Space Administration
NIH	National Institutes of Health
NITRD	Networking and Information Technology Research and Development
NOAA	National Oceanic and Atmospheric Administration
NSTC	National Science and Technology Council
OSTP	Office of Science and Technology Policy
R&D	Research and Development
STRIDES	Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (NIH Initiative)
USGS	U.S. Geological Survey

1. Introduction

Access to advanced computational and data resources has powered many of the recent advances in artificial intelligence (AI), particularly in the area of machine learning. These resources have greatly accelerated advances in all fields of science and engineering by making large archives of data more readily available for research and powering computationally intensive AI-driven analysis and modeling.

Today, a diversity of resource types and Federal approaches exists to support AI research and development (R&D). In terms of advanced computing resources, the Federal Government supports an ecosystem comprising a broad range of computing architectures and capabilities, including high-performance computing, cloud computing, hybrid computing, edge computing, and new computing paradigms. Across these varied advanced computing resources, Federal departments and agencies have taken steps to enhance access supportive of the AI R&D community.

For example, in the case of cloud computing, several Federal departments and agencies, including the National Aeronautics and Space Administration (NASA), National Institutes of Health (NIH), National Oceanic and Atmospheric Administration (NOAA), National Science Foundation, Department of Transportation, and Department of Veterans Affairs, have launched efforts to leverage commercial cloud computing resources to accelerate federally funded AI R&D. This trend, which has emerged in recent years in complement to existing investments in high-performance computing and is responsive to the direction of Executive Order (EO) 13859 on Maintaining American Leadership in AI,[1] is motivated by the on-demand, elastic, and self-serve access to resources at scale afforded by commercial cloud computing platforms. Furthermore, acquiring commercial cloud computing resources can come at a much faster pace than building and deploying local infrastructure, and provides access to continually updated cutting-edge hardware and advanced software stacks. These resources hold the potential of enabling a broader, more diverse research community with access to the cutting edge.

In November 2020, in accordance with EO 13859, the Federal Government's Select Committee on AI issued *Recommendations for Leveraging Cloud Computing Resources for Federally Funded Artificial Intelligence Research and Development*, detailing four overarching recommendations for the Federal Government to advance the use of cloud computing to support AI innovation: (1) launch and support pilot projects to identify and explore the advantages and challenges associated with the use of commercial clouds in conducting federally funded AI research, (2) improve education and training opportunities to help researchers better leverage cloud resources for AI R&D, (3) catalog best practices in identity management and single sign-on strategies to enable more effective use of the variety of commercial cloud resources for AI R&D, and (4) establish and publish best practices for the seamless use of different cloud platforms for AI R&D.

Acting on those recommendations, the National Science and Technology Council's Machine Learning and AI (MLAI) Subcommittee, the operational arm of the Select Committee on AI, led an effort in 2021 to gather common challenges and best practices from early initiatives. This effort involved development of a vision for future Federal use of cloud computing resources to support AI R&D as a component of the federally funded advanced computing ecosystem and across the broad spectrum of Federal agency missions in a manner that embodies responsible stewardship of taxpayer funds.[2] In tandem with this

[1] Released on February 11, 2019, EO 13589 directs the Secretaries of Defense, Commerce, Health and Human Services, and Energy, the Administrator of the National Aeronautics and Space Administration, and the Director of the National Science Foundation to prioritize the allocation of high-performance computing resources for AI-related applications. https://www.govinfo.gov/app/details/DCPD-201900073

[2] A vision and strategic plan for a future advanced computing ecosystem was developed and issued by the National Science and Technology Council's Subcommittee on Future Advanced Computing Ecosystem in November 2020. https://www.nitrd.gov/pubs/Future-Advanced-Computing-Ecosystem-Strategic-Plan-Nov-2020.pdf

work, the MLAI Subcommittee collaborated with the nonprofit organization Internet2 to facilitate a series of dialogues among agency representatives and commercial cloud computing providers to discuss possible pathways to begin to achieve the shared Federal vision for the future cloud environment that effectively supports AI R&D.[3]

In this report, the MLAI Subcommittee summarizes the key findings of the dialogue described above, capturing lessons learned from the activities spearheaded by various agencies to date and highlighting potential opportunities going forward for optimizing Federal use of commercial cloud as a component of broader efforts to further AI R&D that can accelerate scientific discovery and address societal challenges. Importantly, the Federal Government continues to take steps to broaden access to the full complement of advanced computing resources, as evidenced by the work of the Future Advanced Computing Ecosystem (FACE) Subcommittee, which recently published a *FACE Strategic Plan*.[4] However, for the purposes of this report, we focus specifically on the progress Federal departments and agencies are making pursuant to the cloud directive in EO 13859.

Box 1. Federal Initiatives Providing Cloud Computing Resources to Advance AI R&D

Current Federal initiatives leveraging cloud computing resources to advance AI R&D range from time-constrained pilot programs to enduring efforts that have already begun to scale. Examples include the following:

- **NIH STRIDES**: The Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) Initiative is a mechanism that enables the NIH and academic and medical centers to leverage cloud discounts for the conduct of federally funded biomedical research. A partnership between NIH and multiple commercial cloud platform providers, STRIDES supports more than 2,700 academic and medical institutions and more than 300,000 research and research-related individuals, including around 1,200 principal investigators and more than 4,000 postdoctoral fellows in NIH's Intramural Research Program. STRIDES has enabled application of AI-based techniques to a broad range of biomedical research, and is helping unlock the power of data to drive solutions. For example, the open science Serratus program uses a cloud architecture to strive to characterize the planetary diversity of viruses. The effort has identified and made available to the research community tens of thousands of coronavirus and coronavirus-like viral alignments to catalyze a new era of viral discovery—a critical capability for combatting future pandemics.
- **USGS Cloud Hosting Solutions Program**: Focused on facilitating internal U.S. Geological Survey (USGS) research, the USGS Cloud Hosting Solutions Program provides a cloud-based computing and development environment complemented by AI support services to enable the application of AI solutions to priority USGS research efforts. After just a year of operation, the USGS cloud program was able to support 29 AI use cases, both science focused and operations focused. Among those initial cases, four transitioned into production, including a system that detects and predicts water quality sensor malfunctions, and one that automatically identifies bat species present on audio recordings.
- **NSF CloudBank**: The NSF-funded CloudBank Project enables NSF-supported academic researchers to leverage cloud computing to support federally funded research. Through CloudBank, US academic researchers can access multiple commercial cloud platform providers. CloudBank has democratized cloud for a broad range of AI-based computational science and engineering.

---

[3] Internet2 published the results of this dialogue, which took place over the course of three convenings held in August and September 2021, in a white paper released on October 14, 2021: Internet2. 2021. "Research Computing in the Cloud." https://internet2.edu/community/research-engagement/research-computing-cloud-final-report/

[4] See https://www.nitrd.gov/pubs/Future-Advanced-Computing-Ecosystem-Strategic-Plan-Nov-2020.pdf

2. Lessons Learned

The scale and maturity of Federal programs leveraging commercial cloud computing resources to advance AI R&D varies significantly; however, certain patterns and commonalities have emerged with respect to the benefits of these investments, best practices in deployment, and shared challenges. These observations are summarized in the following text.

Benefits of Investments

Agencies that have undertaken early efforts to leverage commercial cloud computing resources to advance AI R&D have commonly experienced benefits to their investments in terms of providing internal and external researchers persistent, on-demand access to cutting-edge capabilities, accelerating experimentation and the use of AI in new domains, and enabling reproducibility and scalability of the research activities and results. Through commercial clouds, agencies have been able to more quickly access specialized AI hardware than would be possible through procurement of on-premise capabilities. Moreover, the commoditization of cloud computing platforms affords constant refreshing of the technology, ensuring agencies have access to the latest computational capabilities afforded through this medium.

In particular, multiple programs were able to facilitate rapid access to advanced computational capabilities, notably graphics processing units (GPUs), to support deep learning and enable collaboration among government and non-government researchers—including both specific projects and hackathons and collaborative coding events. For example, the Air Force Research Laboratory leveraged its cloud computing platform to conduct large-scale AI training experiments with advanced GPUs, which remain very difficult to purchase but could be accessed via the commercial cloud computing platform. NASA hosted hackathons and collaborative coding events such as regular "jam sessions" to educate the NASA workforce on the use of cloud computing platforms to advance its work and to leverage services that have been approved for NASA use.

In addition, use of the cloud has simplified computational access to data owned and maintained by Federal agencies, facilitating efficient use of and collaborative work with big data. For example, 36.4 petabytes of public and controlled access genomic sequencing data hosted by the NIH's National Library of Medicine are now available on two commercial cloud computing platforms,[5] and 10 petabytes of public weather and environmental data are now accessible through the NOAA Open Data Dissemination Program across three commercial cloud computing platforms.[6] NASA has taken similar steps, storing newly collected Earth Science data in the cloud to make it easier for the public to access and reduce the requirement of downloading data to perform analytics.[7]

[5] https://www.ncbi.nlm.nih.gov/sra/docs/sra-cloud/

[6] https://www.noaa.gov/information-technology/open-data-dissemination

[7] https://earthdata.nasa.gov/eosdis/cloud-evolution

Best Practices

Through agencies' experiences, several best practices have emerged around securely and effectively making resources available through commercial cloud instances. Among these, key practices include the following:

- **Dedicated administration teams**. Building this capability has provided agencies with the necessary expertise and authority to manage and oversee access to cloud computing resources, services, and platforms. Such teams have also provided training to the user community and vetted appropriateness of requested resources for achieving specific research goals.
- **User authentication**. Most of the programs have restricted access to known, qualified, and credentialed users. Many also require two-factor authentication as a component of their security measures. Together, these measures provide a baseline level of security and an ability to create user-based access controls.
- **Training and education**. Training assistance and educational opportunities have been critically important for addressing existing skills gaps, advancing equitable access opportunities, and building expertise among the user base. Making these resources available has helped the supported researchers navigate the various cloud computing resource offerings and match specific research and needs to the right compute architectures and software tools.
- **Pre-computed resources and workflows**. Particularly when supporting internal or mission-focused research efforts, pre-computed workflows have reduced duplicative work and created accessible baseline approaches for common starting points for analyses.

Common Challenges

Agency experiences have also helped define a set of common challenges in implementing strategies to enhance access to cloud computing resources. Reflecting the early stages of adoption of cloud computing resources, challenges range from efficiently managing user authentication and access privileges to budget and cost constraints.

Governance and administration of cloud computing programs have presented implementation challenges from multiple angles. For example, early agency efforts have shown how authenticating users can create bottlenecks related to verifying identities and provisioning sign-on capabilities. The underfunding or understaffing of governing organizations can lead to delays in account activations and resolution of issues that arise at every level of access. Furthermore, a lack of authoritative agency and government-wide guidance on approved services, which includes variable data privacy and access considerations, slows adoption and creates variation across agency policies and procedures.

Agencies have also encountered challenges relating to (1) the costs of data storage and access, complicating the ability for multiple teams to access shared data, and (2) ensuring that the users of a given cloud computing platform can locate and maintain awareness of data, experiments, and results relevant to their work and interests. In addition, a persistent question has been determining ways to host and facilitate access to the right kinds of data with appropriate privacy and security safeguards, subject to budget considerations, changing research priorities, and the evolving user community being served. Integration of cloud services with non-cloud resources has also presented challenges in terms of enabling researchers to effectively access the full breadth of agency resources.

Cost, billing, and budget constraints, inherent in any technology investment, are complicated by the variability of cloud computing costs per project and the undesirable ease with which researchers can inadvertently exhaust credits through the use of incorrect settings. Furthermore, variable charges add complications for Federal procurement processes, as does uncertainty around which appropriations categories can be used to purchase which compute capabilities.

Finally, workforce development has remained a critical limiting factor in the ability to adopt and scale cloud computing-based research efforts. Many Federal employees have limited familiarity with cloud computing technologies, and few have industry certification on cloud computing systems. These limitations challenge both internal research efforts and the ability to provide guidance and resources to external researchers.

3. Vision for the Future

A future in which Federal agencies can fully leverage commercial cloud computing platforms to advance AI R&D and drive discovery across areas of science and engineering would be one in which the barriers to entry and to collaboration are lowered, with services tuned to the needs of the research community. It would be characterized by a seamless multi-cloud ecosystem that would enable access to compute, data sets, and storage across multiple commercial cloud computing providers in a manner that maintains security, privacy, and accountability. Facilitated by open-source tools and single sign-on privileges, such a construct would ease adoption and reduce the need for manual processes related to credentialing and collaboration. Usability could be further enhanced through automated data and resource discovery that would unlock knowledge from a dynamically accessible corpus. Approaches such as a cost-effective federated data mesh would help the research community by enabling the movement and replication of data across multiple cloud service providers.

Researchers would have access to training tailored to various levels of use and expertise, opening opportunities for more equitable access to these resources, and would be supported by skilled administrators who would be able to guide and monitor usage.

Finally, adoption would be simplified through authoritative Federal governance and processes that would facilitate responsible and accountable contracting and procurement at various control levels harmonized across agencies as well as a government-wide approach to enabling cloud computing access for external, funded researchers. Agencies would be well served by a Federal approach that keeps close watch on upcoming services and, as appropriate, authorizes those services for broad agency use. Further, a Federal Government-wide approach to enabling access to commercial cloud computing resources for AI R&D would afford researchers with a common user experience regardless of the funding agency, thereby accelerating research progress and making cutting-edge cloud resources more accessible.

Box 2. Internet2-Hosted Dialogue on Research Computing in the Cloud

In the summer of 2021, recognizing that accelerating the adoption of commercial cloud computing is critical to advancing the US research and education community over the coming decade, Internet2—a nonprofit, member-driven advanced technology community—convened a series of discussions among cloud service providers and Federal agencies. These discussions helped the MLAI Subcommittee identify key roadblocks to the broader adoption of commercial cloud computing as well as potential solutions to remedy these challenges. Representatives from the largest providers of cloud computing services—Amazon, Google, IBM, and Microsoft—joined representatives from around 10 Federal agencies in the series of conversations to discuss shared challenges and to individually propose actionable solutions that would serve to advance the US research and education ecosystem. This report, in part, distills the key takeaways and opportunities gleaned from that effort.

4. Opportunities Looking Forward

Achieving this envisioned future will require action on the part of the Federal agencies using cloud services to support AI R&D as well as on the part of cloud service providers.

The Internet2-hosted dialogue provided a chance for the Federal Government to identify actionable opportunities, focusing on the shared challenge areas of financial models, enabling technologies, workforce development, and portals. By considering broad experiences with cloud computing at various agencies, together with what was learned from the Internet2 dialogue, the Subcommittee identified several key opportunity areas.[8]

To address financial models, the Federal Government could do the following:

- **Take better advantage of the purchasing power reflected in the consolidated Federal investments in commercial cloud computing platforms.** This action would facilitate access to the most advanced capabilities of the cloud and provide a means to speak with a common voice on the expectations and needs of the federally funded AI research community.
- **Create explainable models with corresponding costs to better manage budget uncertainty**, as these would illustrate for researchers and program managers the cost dynamics associated with cloud computing, particularly in terms of decisions related to the training parameters and processes.
- **Capture and share best practices from agency cloud programs regarding the contractual agreements and strategies to manage overspending.**

To move toward the envisioned seamless, multi-cloud environment, agencies could do the following:

- **Leverage and help cultivate open-source technologies that can support standard ways of building and executing workloads for multi-cloud deployment** (e.g., containerization and automation).
- **Facilitate and automate identity and access management** through federated systems that bring together the research community inside and outside government.
- **Conduct an evaluation to assess the feasibility of developing a federated data mesh to reduce data movement and replication.**

Agencies could be further assisted in their adoption of commercial cloud computing resources through the following:

- **Creation of a portal kit that would outline a standard template and put forward best practices to implement portals at various levels, depending on the organizational needs.** These approaches, such as standardized approaches for application programming interfaces (APIs) for server and serverless workflows across multiple clouds, could potentially lay the foundation for building managed research workflows as a service on various cloud platforms.
- **Provision of a guide for resource selection** that would help agencies determine the circumstances under which different types of resources are best suited, such as the choice of cloud versus high-performance computing and commercial offerings versus on-premise machines.

[8] The full report from the dialogue can be found in: Internet2. 2021. "Research Computing in the Cloud." https://internet2.edu/community/research-engagement/research-computing-cloud-final-report/

- **Offer a guide on approved policies, procedures, resources, and services** when it comes to commercial cloud offerings, to the extent practicable, by leveraging the purchasing power described above.

Finally, addressing workforce development needs will require the following:

- **Investments in training resources that can serve the entire range of end users, researchers, and technical staff, differentiated for their skill levels, needs, and interests.**
- **Recruitment and retention strategies that include high-demand skill sets that support cloud computing**, such as cloud architects, research computing and data professionals, research software engineers, and data scientists.

5. Next Steps

The vision developed and opportunities identified in this report will serve to inform various initiatives across the Federal research ecosystem, to include efforts to envision a National AI Research Resource and a National Discovery Cloud.

The MLAI Subcommittee will continue to progress toward the proposed vision for the future, including the pursuit of actions related to the opportunities that emerged from the Internet2-facilitated dialogue.

Furthermore, the MLAI Subcommittee will continue to collaborate with other NSTC bodies, notably the FACE Subcommittee, to enhance access to the full and diverse range of advanced computing resources for the AI R&D community.