Opportunities in Artificial Intelligence and Machine Learning

A report of the Cyberinfrastructure Research and Innovation Working Group, Advisory Committee for Cyberinfrastructure

May 2021

Report

Opportunities in Artificial Intelligence (AI) and Machine Learning (ML)

U.S. National Science Foundation Advisory Committee for Cyberinfrastructure[1]
Cyberinfrastructure Research and Innovation Working Group
Tilak Agerwala, Rommie Amaro, Tiziana DiMatteo, Ed Lazowska, Padma Raghavan, Valerio Pascucci, Valerie Taylor

Contents

[1] *The ACCI provides independent advice to the U.S. National Science Foundation (NSF). All opinions, findings, and recommendations expressed within this report are those of the ACCI and do not necessarily reflect the views of the NSF.*

1. Background and Charter

The Advisory Committee for Cyberinfrastructure (ACCI) initiated the Cyberinfrastructure Research and Innovation Working Group (CIRIWG) in response to the Office Director of the Office of Advanced Cyberinfrastructure (OAC).[2] The CIRIWG has its roots in a 2016 National Academies Report on "Future Directions for NSF Advanced Computing Infrastructure to Support US Science in 2017-2020" [1] and built on the 2018 ACCI report "CI2030: Future Advanced Cyberinfrastructure." [2]

The Charter (Revised May 2020) of the working group is: The Cyberinfrastructure Research and Innovation Working Group will identify areas of cyberinfrastructure research, including hardware, software, and middleware, that if executed successfully, yield an integrated Cyberinfrastructure (CI) ecosystem that enables the Nation's researchers to continue to work at frontiers of science and engineering. In these areas, there must be close coupling between those who advance computing technology and those who utilize computing technology to advance science and engineering, creating a "virtuous cycle" in which advances in computing drive advances in science and engineering, which in turn drive further advances in computing. The research areas will be prioritized based on relevance to OAC and the ability to establish virtuous cycles.

Current WG Members: Ed Lazowska, Padma Raghavan, Tilak Agerwala (Chair), Rommie Amaro, Tiziana DiMatteo, Valerie Taylor, and Valerio Pascucci. (Previous WG members: Deborah Dent, Karen Willcox, and Kristin Persson.)

In reviewing areas of cyberinfrastructure research, the working group intentionally focused on opportunities in data science, AI, and ML. The working group concluded that there are significant gaps in the OAC research plan, which can be addressed by new AI and ML initiatives. Research themes addressing these gaps are described below in three broad areas: (1) The "Missing Middle": Data sources to science outcomes and results dissemination; (2) Intelligent (self-managing) CI systems and services; and (3) New models and paradigms for S&E discovery, based on AI. Although this report focuses on data, AI, and ML, the working group recognizes the critical role of mechanistic/physics-based modeling and simulation-based science [3] and the need for continued investment in developing these areas. Indeed, for many of the complex scientific application areas targeted by NSF programs, deep integration of data and models and classical perspectives from applied mathematics and inverse theory will be essential. This report does not discuss the challenges and opportunities at those interfaces.

[2] What constitutes CI research and what is the OAC research agenda? How does OAC evolve its priorities and programs to address the growing need for an integrated CI ecosystem?

2. Findings – Core Research Themes

The “Missing Middle”: Data Sources to Science Outcomes and Results Dissemination

Theme 1: Data-Model Integration

Our first theme—and corresponding recommendation—recognizes that data-centric approaches face limitations since many scientific grand challenges suffer from the lack of adequate sampling of the processes underlying complex, large-scale systems. Also, in many areas of science, data alone cannot provide sufficient insight and explainability to enable true scientific advancement. Data must be combined with existing approaches to address these critical requirements. This is particularly true for a myriad of applications across engineering and geosciences. A great deal is known for many of these systems regarding the underlying physical principles or governing equations; we must continue to appeal to computational science and engineering to unleash this information. While there are many exciting and emerging opportunities in data science, these will be muted without proportionate investments in computational sciences research in data-model integration.

Relevance to OAC: Given the diversity of forms in which scientific data and knowledge are presented in different disciplines and applications, a correspondingly wide-ranging set of methodologies is needed to integrate physical principles and, ideally, ML models for sustained progress in complex science and engineering problems. OAC and its association with the newly established AI Institutes across different science domains should create the ideal environment for developing strategies to combine principles of physics-based modeling, data, and state-of-the-art ML techniques. OAC should also contribute to accelerating the cross-pollination of these ideas among diverse research communities.

Theme 2: Robust Data Harnessing and Domain Driven CI “Prototypes”

Robust Data Harnessing

The role of data has transformed from simply being an “outcome” to being a key driver shaping advanced modern research strategies. A significant challenge across nearly all domains relates to the growing inability to robustly harness the total value of data produced. Data itself has tremendous intrinsic value, but we are limited in tapping into its total value without more effective ways to handle, store, and gain access to the ever-growing body of data across all disciplines.

Data curation, provenance, and dissemination efforts tend to be significantly underfunded and understaffed, as it is usually the “last” thing on a researcher's mind and not rewarded in our scientific merit system. New modalities for sharing data across the scientific community, including community “vetting” of such data, are needed. Enhancing the “findability” of data is a challenge that will also improve data sharing and reuse.

Data production pipelines that include more formal workflow frameworks have an essential role in these efforts, where workflows can also validate similar data from different entities. Enormous volumes of data are produced every day in the scientific community. Still, very little of it adheres to FAIR principles[3], such that it can be robustly reused (for general, as well as AI and ML purposes discussed in Theme 4). This is a significant opportunity area that will benefit from being developed in close collaboration with domain directorates. Challenges include data variety from many experimental and computational techniques, and volume, both of which are rapidly increasing. This includes the more ubiquitous and effective integration of data collected in harsh conditions, where remote access presents additional challenges, as well as sensor data, which needs not only data federation and provenance but also integration into science, engineering, and discovery workflows that run in conjunction with such data coming in real-time.

Significantly increased investment in robust infrastructure and efforts on 1) data *curation*, 2) data *validation*, 3) data *dissemination*, and 4) data *federation* and *provenance* are urgently needed. There is a genuine demand and immediate ongoing opportunity for closer partnerships between domain experts and data science researchers. A closely related aspect and outcome that must be kept in mind throughout is data *access*. Discovery will be further enabled by the ability to search and find data quickly. Something similar to Semantic Scholar, but for the actual data, not just the research publications, is envisioned as a long-term goal.

Increased crosstalk between disciplines regarding lessons learned and best practices for handling data within particular domains would help avoid duplication of efforts. Data handling has been left primarily to domain scientists to “figure out” with relatively little-to-no formal support. This model is not sustainable, nor is it in the best interests of the individual domains. Domain-driven prototyping and domain-independent development are critical - but so is sharing the knowledge and lessons learned across other areas to expedite progress in robust data harnessing.

Crosstalk also must be carried out in strong collaboration with international partners and efforts.

Domain-driven CI Prototypes

Domain-specific prototypes must be developed while noting there will be shared technology among domains that can be reused and eventually linked in some way to build the envisioned semantic scholar type framework for data.

As a concrete example, consider DeepMind's AlphaFold machine learning technology, which showcased a breakthrough success in the recent CASP competition for protein structure prediction [4]. It has been widely acknowledged that the success of DeepMind hinged greatly on the long-term (> 30 years) effort of the World Wide Protein Data Bank, which has robustly provided all five data-related aspects mentioned above for the protein structure data in a reliable, trustworthy manner for the entire

[3] FAIR: Findability, Accessibility, Interoperability, and Reusability. https://www.nature.com/articles/sdata201618

international community. For AI and ML methods to be fully realized—to see successes on par with DeepMind, but in other areas—serious data management efforts are needed across many domains.

Another recent example involves the NSF Molecular Sciences Software Institute's (MolSSI) COVID19 Molecular Structure and Therapeutics Hub.[4] The COVID19 hub was quickly established in response to a real and defined need from the scientific community (COVID19 structure-related related research) to share "final" datasets and all the components required to generate such datasets (e.g., models and methods, intermediate system files, etc.). International collaboration was also a vital component of this effort that has amplified the impact of the activity. Community vetting of submitted datasets touches on the trust aspects discussed in Theme 3. Although the exercise was relatively prototypical and has several areas notable for improvement, it stands as one concrete example of assembling a global, multifaceted community around data sharing and robust data harnessing principles. Creating incentives and opportunities for such efforts will be very important for the robustness of the data science enterprise and increasing consequences in AI and ML activities.

In astrophysics, the National Aeronautics and Space Administration has supported, curated, and archived large datasets drawn from its missions for decades to meet increasing levels of challenges. These centers have gone beyond being simple repositories of raw telescope data. They have become places where high-level data products and data analysis tools are distributed to the science community—often in close proximity to the data. These legacy datasets have proved to be of inestimable scientific value.

Astrophysics has also entered the full sky survey era with even bolder time-domain surveys planned for the next decade—starting with the Large Synoptic Survey Telescope, LSST, or WFIRST—across all wavebands. The immediate need is to develop new application programming interfaces and environments to support advanced machine-learning classification algorithms and methods to efficiently reduce the dimensionality of highly complex, huge datasets and compare them to simulations. This will require quick abstraction and generation of scientific hypotheses from simulations that can then be tested more rigorously in the science cloud or with appropriate machine-selected data subsets.

To address the challenges unique to the given science domain, OAC needs to develop domain-specific CI prototypes that include considering the instruments and analysis frameworks unique to the domain. Further, these prototypes are essential for domain knowledge to be incorporated into the AI and ML methods used with the data analysis. This requires multidisciplinary teams consisting of domain scientists and computer scientists.

Relevance to OAC: If AI and ML are a priority, then robust data harnessing must be a part of that priority. The current scientific grant system encourages but does not enforce good data curation, provenance, and dissemination practices. Every domain is different, but for each domain, some

[4] https://covid.molssi.org

centralized repository—similar to the Protein DataBank—should be considered, which may need federal interagency collaboration and agreements between domain and cyber-infrastructure leadership. Leaving the responsibility of harnessing data to the single PI has not proven successful, nor can we rely on publishers to do this properly. OAC should lead in developing and making available domain-specific CI prototypes for science communities with unique challenges.

Theme 3: Trust and Explainability for AI

As discussed in [5,6], it is crucial to establish trust for AI, which addresses the question of whether an AI model has been constructed, trained, and deployed such that it is appropriate for its intended use; for the case that a model is not suitable it is essential to understand why this is the case. Trust is crucial and depends on how the AI models are used for predictions and decision-making. Trust includes the following topics: robustness, reproducibility, validity, uncertainty quantification, explainability, and interpretability. As noted in [7,8], intuition is also essential for explaining machine learning models, as intuition is critical in human reasoning. The issues of trust and explainability are vital issues to be addressed, especially with the increasing use of AI to accelerate science and engineering advances.

AI provides an opportunity for developing data-driven surrogate models that can be orders of magnitude faster than the simulations of complex functions. As noted in [6], some of the issues with surrogate models include proofs of interpolation/extrapolation, robustness, and assessment of confidence associated with the model predictions. Further, it is essential to quantify the uncertainties concerning the development and application of surrogate models.

Many physical systems, such as weather forecasting or climate modeling, include data from multiple modes -- such as sensors, simulations, and observational data from instruments. Further, data can be streamed, requiring real-time analysis that can be used for feedback and control. For example, using AI with edge devices connected with instruments can detect anomalies and provide data to actuators that can be used to control instrument settings to address the anomalies. The different modes of data have other characteristics and formats, resulting in additional challenges related to data fusion to establish correlations across the different modes.

Relevance to OAC: With the increasing use of AI and ML to accelerate scientific advancement, it is essential to establish trust and explainability in knowing that an AI model has been validated and is appropriate for its intended use.

Theme 4: Machine Learning Across the Cyber Infrastructure (MLCI)

As noted in [2], the growth in instrumentation capabilities results in rapidly growing data sets, increasing the challenges of sharing, analyzing, modeling, and visualizing the data quickly and efficiently. Taking computing and storage to the data source, or edge computing, coupled with AI and ML methods at the edge, provides a way for addressing these challenges. Edge computing allows data collection and analysis at scales and rates not otherwise possible, data-based instrument steering, and facility management. Robust and versatile edge computing devices enable scientists to perform large-scale science experiments in harsh environments (e.g., Arctic Sciences research programs in the Office of Polar Programs). Edge computing, coupled with actuators and robotics, provides opportunities for real-time steering of instruments based upon the data, empowering scientists to modify experiments in real-time. Particle accelerators, light sources, and complex instruments have many control points and require high stability levels. Distributed ML-based edge computing can help automate and optimize operations. This is a challenging problem due to the lack of prior models. Edge computing is particularly relevant to NSF's Large Facilities.

Unique features of edge computing, such as real-time, parallel, and distributed processing, result in new challenges in data security and privacy-preserving algorithms: (1) Lightweight data encryption methods and fine-grained data sharing systems based on multiple authorized parties; (2) Distributed access control (3) Efficient privacy-preserving schemes.

Machine Learning has a tremendous potential to accelerate discovery across all areas of science and engineering. ML will permeate the entire computing continuum from instruments to sensors/actuators, edge computing, clouds, and traditional HPC systems [6].

R&D on processor and system architectures for ML for science is restricted mainly to industry and national labs. OAC should enable the research community to explore and assess the new MLCI systems' efficacy in specific domains and their potential for transforming scientific research. Testbeds (including public clouds, edge computing, HPC systems, and quantum computing) facilitate the CI ecosystem's use by domain scientists and provide the infrastructure needed to explore processor and systems architectures and develop the required full software stack.

Relevance to OAC: OAC is the leading provider of computational resources to the scientific and engineering research community. Consistent with the CI 2030 [2] recommendation, this role must continue, but the approach must broaden to include the entire computing continuum. MLCI testbeds, MLCI architecture, and software research are immediate requirements with impact within five years.

Intelligent (Self-Managing) CI Systems and Services

Theme 5: Deeply Network-Aware, Reactive, and Proactive Scientific Workflow Management Systems and Software-Defined Cyberinfrastructure and Intelligent Networks

The introduction of AI/ML technologies within scientific workflows with a deep awareness of the underlying network infrastructure can enable self-optimizing, self-healing workflows that take full advantage of the available resources and use system logs and performance metrics to achieve optimizations that traditional approaches would achieve only in a long time and with a high degree of human specialization. This area is in high demand for CI-specific research and deployment since it must be tested on the specific conditions of the current and future NSF-funded cyberinfrastructure. If successful, it is positioned to give massive benefits to a broad set of scientific activities.

Software-Defined Cyberinfrastructure and Intelligent Networks

As observed in Reference Number 235 in [2] to reach the scales implied by the accelerator and detector upgrades, the cyberinfrastructure itself needs to become more intelligent and capable, assuming more responsibility local to the resource. A software-defined cyberinfrastructure effectively decouples the physical hardware and the operating system's low-level services from how users “view” the resources available to them. Instead, a set of rules allows implementing a particular transaction (such as moving data between specific storage components) and presents the user with a higher abstraction level. Automated, machine-learning-based decision engines coupled with a software-defined infrastructure are a viable path towards dynamically optimizing globally distributed resources. Emphasis should be given to the construction of intelligent networks that can yield a relatively simple but powerful connectivity infrastructure. Intelligent networks introduce the possibility of offering a broad spectrum of services that support a distributed environment involving several classes of applications with very different requirements, including bulk data transfers, interactive sessions, batch processing, real-time services, etc. The intelligent layer can gather and harmonize all the requirements and present a compatible view with their needs without modifying their application and with a minimal performance penalty. The set of services provided by the NSF-funded cyberinfrastructure is increasing in variety. Traditional scheduling of resources –especially network and storage—is inadequate to address the users' needs and the CI professionals managing the facilities. Future research and development may lead to the unification and generalization of software-defined cyberinfrastructure resources enhanced by AI and ML management.

Relevance to OAC: Providing a flexible and intelligent cyberinfrastructure is a core responsibility of OAC. Intelligent management of cyberinfrastructure resources is crucial for future workloads and CI management

New Models and Paradigms for S&E Discovery Based on AI

Theme 6: New Models and Paradigms for S&E Discovery Based on AI

Artificial Intelligence, machine learning, computer vision, natural language processing, robotics, and more will revolutionize scientific discovery just as it is revolutionizing so many other aspects of our lives. As an example, consider the impact that DeepMind's AlphaFold machine learning technology has had on the protein folding problem, where it showcased dramatic improvements in the ability to predict protein structures—even outside of known systems==in the recent CASP13 competition [4].

In general, computer scientists, applied mathematicians, and statisticians driving advances in machine learning left to their own devices are unlikely to focus on scientific discovery as an application area to guide their research. But the opportunities and the challenges of scientific machine learning are extraordinary— both in numerical and data-intensive domains.

Basic research needs scientific machine learning for both domains—numerical and data-intensive— are well described in [3]. This document is an excellent reference for the specifics of what is required.

The applications of ML to data-intensive discovery are amply evident and sit adjacent to commercial applications of ML—although several unique challenges arise. As described in [2] and in Themes 1 and 2, we need new frameworks and tools for data-intensive scientific discovery: computational abstractions for research domains, coupled with methods and tools for their analysis, synthesis, simulation, visualization, sharing, and integration; standard tools, approaches, and frameworks for handling big data across different disciplines; new tools that incorporate insights from emerging statistical analysis and machine learning approaches, among others; and solutions for end-to-end management of big data, from ingesting massive amounts of data to archiving the products of data analyses. Sparse data represents a significant challenge. Earth observation, for example, continues to be a sparse data system; we wind up training neural networks using simulations because we lack data. There are fields in the physical sciences where equations – laws – provide constraints, and unconstrained machine learning is not useful.

Applications to modeling, simulation, and decision support are more unique to the process of discovery; see, for example, Priority Research Direction 5, focused on machine learning-enhanced modeling and simulation, and Priority Research Direction 6, focused on intelligent automation and decision support, in [5]. In modeling and simulation, human expertise is typically integral in the simulation process for performance, robustness, and fidelity; there are tremendous gains to be realized through ML. In the realm of decision support, ML can assist with how to evaluate a complex, expensive simulation model over a high-dimensional parameter space, how to combine experimental and simulation data best to inform decisions, and how to validate the resulting evaluations and translate their uncertainty into quantifiable confidence for a decision-maker. Managing the interplay between automation and human decision-making is critical. Of particular interest are system-level, mechanistic, computational models of physical, biological, cognitive, and social systems that enable integrating different processes into coherent and rigorous representations that can be analyzed, simulated, integrated, shared, validated against experimental data, and used to guide experimental investigations. These models must cross the levels of abstraction and disciplinary boundaries to allow studies of complex interactions – e.g., those that couple food, energy, water, environment, and people.

Considering the future improvements in the sensitivity of Gravitational Wave (GW) detectors and their ability to detect many events per week, ML techniques are poised to become essential tools in GW science and multi-messenger astrophysics. Current examples of development in this realm include techniques for improving the sensitivity of Advanced Laser Interferometer GW Observatory and Advanced Virgo GW searches, methods for fast measurements of the astrophysical parameters of GW sources, and algorithms for reduction and characterization of non-astrophysical detector noise. These applications demonstrate how machine learning techniques may be harnessed to enhance the science that is possible with current and future GW detectors.

As described in Reference Number 279 in [2], cognitive tools for scientists are also a pressing need. The next-generation cyberinfrastructure for science needs to provide a broad range of computational tools that leverage and extend human intellect and partner with humans on a wider range of tasks that make up a scientific workflow (formulating a question, designing, prioritizing, and executing experiments designed to answer the question, drawing inferences, and evaluating the results, and formulating new questions, in a closed-loop fashion).

Relevance to OAC: New models and paradigms for discovery are required to enable cross-domain, convergent research, and improved research productivity. OAC has a long history of coupling researchers advancing computational technology with researchers utilizing computational technology to advance other fields of discovery, creating a "virtuous cycle" in which advances in one field stimulate advances in another. As noted at the start of this section, computer scientists and statisticians driving advances in AI and machine learning, left to their own devices, are unlikely to focus on scientific discovery as an application area to guide their research. At the same time, researchers in application areas are unlikely to be at the forefront of AI and machine learning advances. OAC must provide the bridge.

3. "Virtuous Cycles" and Scale

Discovery in science and engineering has always placed unique—and often extreme—demands on computing hardware and software technologies. But because discovery represents a relatively small "market share" (in dollars, in visibility, etc.), it tends not to be the natural focus of computer scientists, computer engineers, statisticians, and others dedicated to advancing computing technology. At the same time, researchers who rely on computing technology to drive advances in their field are unlikely to be at the forefront of computing advances, particularly in emerging areas such as AI and ML.

The federal government—particularly the Department of Energy and NSF—has played a crucial role in bridging those who advance computing technology and those who employ computing technology to advance discovery in other fields of science and engineering. The goal—the imperative—is to create a "virtuous cycle" in which advances in computing drive advances in science and engineering, which in turn drive and inform further advances in computing.

Innovation in AI has exploded in recent years, thanks largely to the advent of "deep learning." Bringing this innovation to bear on discovery is essential to advancing science and engineering and vital to our Nation's competitiveness and leadership across a broad range of important fields.

OAC should partner with the Directorate for Computer & Information Science & Engineering (CISE) and the other NSF Directorates to partner researchers in AI, ML, and data science with researchers in science and engineering disciplines. These partnerships, at scale, have not been a strength of NSF. For example, the "virtuous cycle" in data-intensive discovery was catalyzed by awards from the Gordon and Betty Moore Foundation and the Alfred P. Sloan Foundation after NSF failed to fund multiple cross-disciplinary proposals in this area. Decades previously, the Whittaker Foundation stepped in to create the field of bioengineering.

4. Recommendations

Investments in the six high-priority research themes below will fill gaps in the OAC research program. They should be made in collaboration with CISE and the various NSF Directorates to partner researchers in AI, ML, and data science with researchers in science and engineering disciplines. The goal is to establish “virtuous cycles" where advances in computing drive advances in science and engineering, which in turn drive and inform further advances in computing.

Advance the development of predictive models and data-model integration. Predictive models and data-model integration remain foundational tools across many science, engineering, and medicine areas. Research in areas such as advanced data assimilation, uncertainty quantification, multi-scale multi-physics approaches, and adapting existing state-of-the-art algorithms will enable scientists and engineers to fully leverage computational science, data science, AI, and ML.

Robust Data Harnessing and Domain-Specific CI “Prototypes.” (a) Develop robust methods of data handling covering all necessary aspects, (b) Support the development of more domain-specific data repositories that model the tremendous success of efforts such as the Protein Data Bank, (c) Develop domain-specific CI prototypes that can be replicated to other domains with adjustments. (d) Help catalyze cross-disciplinary discussions on data harnessing and within domains, but in the broadest sense, with international partnerships.

Trust and Explainability for AI. Partner with CISE and the Directorate for Mathematical & Physical Sciences (MPS) to develop new techniques and methods for the assurance of AI models used to accelerate scientific discovery. Also, it is crucial to partner with the science and engineering directorates that can leverage these new techniques to accelerate advances.

Machine Learning Across the Cyber Infrastructure for S&E. To address future facilities challenges with edge computing and enable MLCI architecture and software research, OAC should consider all modalities to deliver MLCI testbeds to the research community, including partnering (with other government agencies and industry), capitalization, and renting, and continue this practice as new, more capable systems architectures for AI-based discovery emerge.

Deeply Network-aware Workflow Management Systems and **Software-Defined Cyberinfrastructure and Intelligent Networks.** Foster the development of Intelligent, self-managed CI subsystems by supporting partnerships among cyberinfrastructure facilities and professionals, applied AI and ML experts, and application domain scientists. Requirements from CI technology and application workflows need to drive the use of AI and ML techniques via multidisciplinary teams developing community testbeds. Extended time awards will be ideal for making new, experimental resources available to the communities of interest and share insights and solutions.

New Models and Paradigms for S&E Discovery Based on AI. Partner researchers in AI, ML, and/or data science with researchers in science and engineering disciplines to adapt forefront approaches in AI, ML, and/or data science to the requirements of scientific and engineering discovery. These should be large-scale awards in size and duration to attract top researchers' attention in methodology and application.

5. Suggested Connections and Touchpoints between OAC and NSF

The working group identified connections and touchpoints between OAC and other NSF directorates to improve OAC's effectiveness. These efforts are considered essential but are of lower priority than the themes mentioned above.

(a) Modifying Workflows at the Application Level for Data-Centric and Distributed Operation

Currently, there is a focus on providing the HPC community with exascale systems by 2021-2022. Simultaneously, we see the emergence of AI accelerators, cloud systems, and edge computing. It is expected that future CI resources will be distributed and heterogeneous. To facilitate the efficient use of such systems, it is essential to provide workflows that allow science applications to use the given resources to manage and analyze data easily. The data can be generated from sensors, instruments, experiments, or simulations. It is important to provide workflow frameworks that are seamless to domain scientists.

Relevance to OAC: OAC should take the lead in making available frameworks and workflows for the data-centric and distributed operation of science applications. This is in line with the recommendation for NSF to support foundational cyberinfrastructure research for data science, focusing on frameworks, workflows, and tools.

(b) Programming and Debugging New AI and Analytics-Driven Workloads

New AI and analytics-driven workloads present fundamentally new characteristics and require new programming models and debugging tools. For example, debugging an AI workload can lead to several considerations that range from dynamically changing data to the direct influence that a particular set of inputs has on the output or the training of a specific model. This activity probably needs some independent basic research and some research focused on the use within a broader cyberinfrastructure and can substantially impact the way the cyberinfrastructure is managed. The use within the cyberinfrastructure is undoubtedly in the purview of the OAC in conjunction with CNS and CCF.

(c) Machine Learning and Cyber-Physical System Security

As noted in [2], cybersecurity is critical to the integrity of the global computing system. New infrastructure elements are needed to support persistent data collections and researcher identities across multiple collaborations. It is important to have a framework and infrastructure for real-time analysis of multimedia data (voice and gesture recognition, biometrics, sentiment analysis, and social relation graph analytics) to better detect security threats to researcher identity. Data is the major component by which machine learning methods are used to develop models that aid in driving science discovery. It is critical to have a framework to maintain the integrity of data collections and researcher identity.

The end-to-end infrastructure for science is rapidly becoming a cyber-physical system with sensors and actuators that interface to large facilities and other physical experimental apparatus with automated inference capabilities in control and analysis loops, self-driving laboratories, and scientific workflows []. New AI-based cyber-physical security methods need to be developed.

Relevance to OAC: OAC should take the lead on developing secure frameworks via programs such as CICI (Cybersecurity Innovation for Cyberinfrastructure). NSFCISE has an active research program on cyber-physical security. OAC should ensure that the cyber-physical infrastructure for science outlined above is prioritized in the NSF CPS program.

(d) Programming Models for Quantum Computing Where it can be a Game Changer, e.g., Quantum Mechanics Simulations

Physics and chemistry – quantum mechanics simulations – are the “killer apps” for quantum computers, the applications most likely to yield a high payoff in the relatively near term. With useful quantum computers clearly on the horizon, there is a pressing need for algorithms and programming models that address these fields' needs.

Relevance to OAC: Quantum architectures are being pioneered in the industry (IBM, Microsoft, Google, Amazon, and others) by other research agencies and different NSF divisions. It is recognized that NSF has made some significant investments in this area via the Quantum Leap Challenge Institutes (QLCI) program. OAC, via NSF Expeditions, is best positioned to bring together physicists, chemists, and computer scientists to pioneer improved algorithms and programming models for addressing grand challenge problems on quantum architectures.

6. References

[1] National Academies of Sciences, Engineering, and Medicine. 2016. *Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science and Engineering in 2017-2020*. Washington, DC: The National Academies Press. **https://doi.org/10.17226/21886**.

[2] NSF Advisory Committee for Cyberinfrastructure (ACCI). 2018. *CI2030: Future Advanced Cyberinfrastructure.* **https://www.nsf.gov/cise/oac/ci2030/ACCI_CI2030Report_Approved_Pub.pdf**

[3] Rüde, U., Willcox, K., McInnes, L. C., & Sterck, H. D. 2018. *Research and Education in Computational Science and Engineering*. SIAM Review, 60(3), 707–754. **http://doi.org/10.1137/16M1096840**

[4] Senior, A., Jumper, J., Hassabis, D., & Kohli, P. 2020. *AlphaFold: Using AI for scientific discovery*. **https://deepmind.com/blog/article/AlphaFold-Using-AI-for-scientific-discovery**

[5] Baker, N., Alexander, F., Bremer, T., Hagberg, A., Kevrekidis, Y., Najm, H., Parashar, M., Patra, A., Sethian, J., Wild, S., Willcox, K., & Lee, S. 2019. *Workshop Report on Basic Research Needs for Scientific Machine Learning: Core Technologies for Artificial Intelligence. United States*. **https://doi.org/10.2172/1478744**

[6] Stevens, Rick, Taylor, Valerie, Nichols, Jeff, Maccabe, Arthur Barney, Yelick, Katherine, & Brown, David. 2020. *AI for Science*. United States. **https://doi.org/10.2172/1604756**

[7] Selbst, A.D. and Barocas, S., 2018. *The intuitive appeal of explainable machines*. Fordham Law Review, 87, p.1085. **https://ir.lawnet.fordham.edu/flr/vol87/iss3/11**

[8] Lipton, Zachary C. 2018. *The Mythos of Model Interpretability*. Queueing Systems. Theory and Applications 16 (3): 31–57. **https://doi.org/10.1145/3236386.3241340**