Report from the NSF Workshop on Smart Cyberinfrastructure 2020

Alexandria, VA, USA, February 25-27, 2020

This workshop was funded under NSF grant 1941085

NSF Workshop on Smart Cyberinfrastructure 2020

Crystal City, Arlington, VA, USA, February 25-27, 2020

http://smartci2020.org

Table of Contents

Executive Summary

Machine learning and other Artifical Intelligenece technologies (all indicated in the following as AI) used within a modern, smart cyberinfrastructure have become critical new avenues for discovery and validation in data-driven science and engineering disciplines of all kinds. We can expect many landmark discoveries and new lines of productive research to be enabled through AI analysis of the rapidly growing treasure trove of scientific data. AI-based techniques have been applied in many fields of science and engineering, including remote sensing, cosmology, energy, cancer research, IT systems management, and machine design and control, but the lack of proper integration with the current NSF-supported cyberinfrastructure is limiting their potential. Recent events due to the COVID-19 pandemic have highlighted how cyberinfrastructure is a crucial enabler of modern research, with massive simulations and data management capabilities [8-10], but these events have also emphasized how the lack of proper integration with AI technology remains a major limiting factor for the advancement of science and engineering, especially when any kind of rapid response is needed.

Summary of the Workshop

The 2020 NSF Workshop on Smart Cyberinfrastructure was held at the Hyatt Regency Crystal City at Reagan National Airport in Arlington, VA, on February 25-27, 2020. A steering committee of domain experts and CI community representatives led the organization of the workshop. Workshop attendance was by invitation only and consisted of 60 participants, including AI/ML experts, domain scientists and engineers, representatives from the National Science Foundation (NSF), CyberInfrastructure (CI) professionals, and representatives from NSF-funded computing facilities.

The overarching goal of the workshop was to develop a community view of the state of the art and research needed in the use of scientific cyberinfrastructure to advance AI technologies as well as how to use these technologies to manage NSF-sponsored cyberinfrastructure efficiently. The scale and high priority of AI research activities and opportunities call for the urgent and coherent development of a smart cyberinfrastructure that enables accelerated progress, shared outcomes, workforce training, and operational excellence. Developing a shared research vision allows us to identify the most crucial gaps in existing capabilities that are least likely to be filled by industry or existing efforts by NSF or other agencies.

The pre-workshop activities involved an assessment of the state of the art via white papers that were collected and summarized by the steering committee. The 35 responses with white paper submissions were shared with the attendees prior to the workshop and made available to the general public through the workshop website. The workshop agenda was designed to maximize the interactions among the participants with 6 active breakout sessions and 4 panel discussions, in addition to questions and feedback during 3 keynote and 9 plenary presentations as well as 13

lightning talks. Based on the content of the white papers, the discussions were primarily focused on the following 4 themes:

1. exploring and assessing the cyberinfrastructure requirements for AI/ML-based scientific applications,
2. accelerating AI/ML algorithms on NSF-enabled cyberinfrastructure,
3. promoting the efficient use of NSF-enabled cyberinfrastructure through AI/ML technologies,
4. enabling easy and productive access to AI/ML tools by a large community of domestic and international science stakeholders.

Overall, the enthusiastic engagement of all participants made the workshop a successful forum for open discussions, with a broad spectrum of perspectives from all stakeholders. This report documents and organizes the wealth of information provided by the participants before, during, and after the workshop. Additional details can be found at http://smartci2020.org/.

Working Groups

From the initial 4 areas of discussions, the workshop activities led to the creation of the following 6 working groups, which collected and summarized the community feedback and distilled the associated findings and recommendations reported later in the report:

- **Cyberinfrastructure Requirement for AI-Based Scientific Applications:** This working group explored the requirements for cyberinfrastructure to facilitate the application of AI in science and engineering applications. As the current CI has become an indispensable tool for modern scientific discovery, a smarter CI is expected to become an indispensable accelerator or gap-bridger for solving the pressing challenges in science and engineering. This working group assessed S&E requirements to achieve this goal, including the availability of hardware components, the effectiveness of the software infrastructures, and the usability of the data repositories that enable exploration of scientific and engineering opportunities, and solutions.

- **Development of the Next-Generation System Cyberinfrastructure Tightly Integrated with AI:** This working group explored how AI can inform the design and operation of NSF cyberinfrastructure, both the hardware and software, in centralized facilities and at the edge. This discussion was facilitated throughout the workshop by a panel, several invited talks, and two breakout sessions. The first breakout session focused primarily on the physical aspects of the "central" CI, inside the datacenter, with the second breakout focused more on "edge" issues, data acquisition, and software layers.

- **Improving AI/ML Associated Software Tools on NSF-Enabled Cyberinfrastructure:** Although we have examples of resources and scientific applications that have successfully deployed artificial intelligence techniques as part of the solution, we face a number of

challenges in scaling and generalizing the developed methods and tools at the CI level. Thus, this working group discussed the state of the art in this area and explored the potential improvements needed in AI/ML algorithms, software infrastructure, and data repositories to enable effective use of these technologies for (future) NSF infrastructure.

- **Improving AI/ML Associated Data Repositories on NSF-Enabled Cyberinfrastructure:** This working group discussed the opportunities for building shared data and modeling repositories for training and inference, with applications in computer science, collaborative problem solving, and evidence-based CI development. These shared datasets include repositories of scientific data generated on the NSF-supported CI as well as the large amounts of data collected while monitoring scientific computations that can be used to enable AI control and optimization of the operations. In both cases, we aim to develop standards and repositories that facilitate data sharing and retargeting the information available for new AI tasks.
- **Human Capacity Development in AI and Smart Cyberinfrastructure:** This working group gathered community feedback regarding the needs and requirements for the CI workforce (intended in a broad sense), comprised of all participants and contributors to the CI ecosystem, including students, staff, research professionals, scientists, and engineers at every level of expertise. This "human capacity" is the strategic asset that will allow us to ultimately deploy the smart cyberinfrastructure of the future by creating new knowledge, assess best AI practices, support end-users needing AI tools, and create/participate in training opportunities.
- **International Collaborations:** This working group discussed existing challenges and proposed solutions associated with the ability to leverage international collaboration for the development, use, and sharing of AI technologies for a smart cyberinfrastructure.

Key Findings and Recommended Actions

The following findings and recommendations were developed based on information collected during the process, including a survey of the information provided in the white papers before the workshop; an account of the community feedback during the panel discussion and breakout sessions; and content from the plenary, keynote, and lightning talks, including the group discussions that followed each presentation.

Key findings:

1. The usability of AI techniques at all levels of the cyberinfrastructure remains a concern for domain scientists and engineers, ranging from the integration of new hardware to the accessibility of software stacks. In addition, facilitating the interpretation of results, with a proper assessment of uncertainties and reliability, is a major gap within the current CI.

2. A largely untapped opportunity exists for developing new AI technology to improve CI by leveraging the existing logs, workflows, statistics of data collected, and curated data repositories that have been unavailable because they have not been built around data-sharing values.

3. A major CI challenge is the lack of critical mass in the human capacity that creates and maintains scientific and institutional knowledge, and that supports the development and deployment of smart cyberinfrastructure. The lack of educational and training opportunities will make this problem worse.

4. We have a critical lack of hardware, software tools, and systems in support of performant science and engineering application workflows that integrate AI technologies.

5. Institutional coordination among CI professionals, AI innovators, and domain scientists is critically needed to achieve fast technology creation and transfer, making the smart CI a key tool for scientific innovation.

Recommended actions:

1. Foster collaboration between the domain scientists/engineers and CI experts through synergistic funding mechanisms, cross-disciplinary exchange during human capacity development, and AI-augmented science/engineering curriculum.

2. Promote the development of architectures, software tools, and systems supporting S&E+AI/ML application workflows, including providing vertically integrated, performant, and portable solution environments to support the responsible use of AI techniques.

3. Enable collaborative research within the CI community through open access to data collected and created from NSF-supported cyberinfrastructure, including the full spectrum of performance monitoring, software usage, AI models, and scientific outcomes.

4. Support the creation of testbeds that facilitate the experimentation of new AI technologies without disrupting CI operations in production environments. In particular, CI testbeds should enable innovation in AI-driven discovery and simulation, managing and decreasing energy

consumption, enhancing smart sensors at the edge, and facilitating the development of advanced software tools integrated with AI technologies.

5. Establish Centers of Excellence (for example, following the model of the Center for Trustworthy Scientific Cyberinfrastructure, CTSC) that connects domain scientists and engineers with the latest AI technology and facilitates its effective use on a smart CI with the adoption of best practices. A Center of Excellence would also constitute a model for AI-CI workforce development and training, identifying career paths, diversity, and international collaborations.

The overall consensus was that the success of the workshop was due to the timely identification of an area of urgent need. The insights gained will need long-term sustained activities that continue to engage and grow this community. In particular, we have initiated a number of discussions in more focused areas to tackle more effectively and, in a coordinated way, the numerous challenges that we will face to build the smart cyberinfrastructure of the future.

1. Introduction

Overview and Goals

A broadly representative and diverse group of 60 scientists, researchers, and engineers met February 25-27 in Crystal City, Arlington, VA to consider the challenges and opportunities being presented by the incredibly rapid developments in AI (including machine learning and other smart technologies) with respect to their use on the NSF-supported cyberinfrastructure. These technologies are of top national importance, as represented by the February 11, 2019, Executive Order on Maintaining American Leadership in Artificial Intelligence [7]. The 2.5-day workshop format was chosen to give full consideration to this important topic and allow active contributions by a broad spectrum of stakeholders, including junior researchers and underrepresented groups.

This workshop took stock of the current use of artificial intelligence in the national cyberinfrastructure and yielded findings and recommendations to further this critical national priority. In particular, a key goal was to develop a common understanding of the current and future requirements for scientists and engineers; assess gaps in the architectures, best practices, and enabling technologies; and identify the needs for creating a critical mass in the human capacity to support the operations, provide training opportunities, and ultimately create the knowledge necessary to create and manage a sustainable smart cyberinfrastructure.

Specific goals of the workshop included:

- Understand the best practices in integrating AI tools in workflows for science and engineering.
- Identify the requirements for incorporating AI tools, quickly and effectively, in new science and engineering applications.
- Identify gaps in the hardware infrastructure needed for the development of a smart cyberinfrastructure from the edge to the leadership-class computing facilities.
- Understand the differences in priorities among industry, academia, and national laboratories to determine the unique role that can be undertaken by the National Science Foundation.
- Develop interoperability guidelines, mechanisms, and processes that can make AI software quickly understood and adopted on the cyberinfrastructure.
- Assess the need for a new class of data repositories that can enable CI-specific research goals that can be addressed with AI tools.
- Understand the AI-specific needs of the CI workforce for deploying a new smart cyberinfrastructure, supporting the science and engineering challenges of NSF priorities, and creating the critical mass in human capacity that can be sustained in the long term.

- Provide recommendations that can serve as inputs to current and future NSF AI-CI related programs.

Workshop Organization

Steering Committee

The steering committee of the workshop had the main responsibility to organize the workshop, identify a proper venue, invite participants with a broad spectrum of expertise, create a call to the community to provide pre-workshop input via white papers, and establish the workshop agenda, including a plan for keynote, plenary, and lightning talks; panel discussions; and breakout sessions. After the workshop, the steering committee was also responsible for producing the workshop report with the added help of 5 supporting editors who joined the group during the meeting. The committee was composed of 9 leading experts in the field representing the CI community and applied use of AI technologies.

The committee members were as follows (bios are included in Appendix C):

- Ilkay Altintas - San Diego Supercomputer Center
- Jose Fortes - University of Florida
- Ian Foster - The University of Chicago and Argonne National Laboratory
- Helen Gu - North Carolina State University
- Salim Hariri - University of Arizona
- Valerio Pascucci (PI and Chair) - University of Utah
- Dan Stanzione - University of Texas at Austin
- Michela Taufer - University of Tennessee Knoxville
- Xinyu Zhao - University of Connecticut

The supporting editors were as follows:

- Peer-Timo Bremer - Lawrence Livermore National Laboratory
- Terry Moore - University of Tennessee Knoxville
- Nick Nystrom - Pittsburgh Supercomputing Center
- Steve Petruzza - University of Utah
- Glenn Ricart - US Ignite

Workshop Attendees

Workshop attendance was by invitation only, based on a list of invitees developed by the steering committee. The workshop attendees included representatives from the CI community, AI experts and practitioners, and domain scientists and engineers. Moreover, the attendees were affiliated with diverse institution types, including academia, industry (e.g., Google), DOE national laboratories (e.g., ANL), international partners (e.g., AIST in Japan), and the National Science Foundation. 60 participants attended the workshop (of 65 registered). Travel support was provided

to a few early career researchers and international partners to broaden the workshop participation. The final attendee list is provided in Appendix A.

Workshop Structure and Activities

Pre-workshop Activities

The pre-workshop activities were divided in two main tasks: (i) the steering committee created four preliminary working groups to compile an initial assessment of the state of the art and open challenges to be presented at the workshop; and (ii) the committee developed a call for white papers to provide additional input on the topics analyzed by the working groups and to guide the selection of additional speakers for lightning talks.

White papers. The attendees who tentatively accepted the workshop invitation were requested to submit a short (about two pages in length) white paper based on their perspectives on at least one of the following suggested areas of interest:

- Accelerate AI/ML algorithms on NSF-supported cyberinfrastructure
- Efficient use NSF-supported cyberinfrastructure through AI/ML technologies
- Provide easy and productive access to AI/ML tools by a broader and diverse scientific community of domestic and international stakeholders
- Explore and assess cyberinfrastructure requirements for AI/ML-based scientific applications

These areas were recommended to help structure the discussions, although one role of the white papers was also to assess if the attendees would highlight other topics of interest that may have been missed by the steering committee. To further facilitate combining the responses, the attendees were provided with the following questions:

- What are the main open research challenges and/or promising research directions in AI/ML and other smart technologies that can positively impact future NSF-supported cyberinfrastructure?
- What computational testbeds for community research can accelerate innovation and lead to future deployments?
- What innovations are needed in the cyberinfrastructure of the future?
- How can the community promote and implement workforce development and training to facilitate the support of cyberinfrastructure?
- What tools and techniques are needed to support cyberinfrastructure-enabled science and engineering?

31

We received 31 white papers, representing a broad set of ideas from the community that were made public and are reported in Appendix E.

Preliminary Working Groups. The steering committee divided the work into the following four initial working groups and appointed one committee member to lead each group:

- Cyberinfrastructure technology in support of AI and smart computing (including data and deployment): Led by **Ilkay Altintas**
- AI in support of a smart cyberinfrastructure: Led by **Dan Stanzione**
- AI in support of science: Led by **Xinyu Zhao**
- Workforce development and international initiatives: Led by **Michela Taufer**

The white papers were analyzed by the members of the working group with three main tasks in mind: (i) analyze the content and prepare to provide a summary to the community at the opening of the workshop, (ii) select a diversity of speakers for lightning talks, and (iii) determine the topics to be addressed during the panel discussions and breakout sessions. After this activity and as a consequence of the discussions during the workshop, the working groups were reorganized as indicated in the executive summary. This revised organization is also used for the content in the remainder of this report.

Workshop Website: The steering committee also created a website dedicated to the workshop activities, which can be found at http://smartci2020.org/. The website was used primarily to communicate to the attendees the details and objectives of the workshop, provide the structure of the agenda, announce the call for white papers, and make immediately available the white papers contributed. We also experimented with a set of calling cards that the attendees were requested to fill out to provide a short summary of their interests and an image to make them more easily recognizable. These calling cards were put on the website before the workshop started and played in a loop at the registration table and on the main screen during the breaks. The feature was highly appreciated, especially by the junior participants, because it made it much easier for everyone to identify one another from the beginning of the workshop and then interact with each other.

Timeline for Pre-Workshop Activities: The pre-workshop activities started during fall of 2019 and continued until the workshop took place. White paper submissions were left open until early February for full consideration, although a few were accepted closer to the date of the workshop. The committee created four presentations that summarized the input from the community and used them to start the technical program with a set of open questions and thought-provoking discussion points that immediately created an atmosphere of intellectual debate with highly dynamic interactions.

The Workshop

The 2020 NSF Workshop on Smart Cyberinfrastructure was held at the Hyatt Regency Crystal City at Reagan National Airport in Arlington, VA, on February 25-27, 2020. Workshop participation was by invitation only and involved 60 participants, including AI/ML experts and practitioners, domain scientists and engineers, representatives from the National Science Foundation (NSF), CI professionals, and representatives from NSF-funded computing facilities. The program of the workshop, which was based on the white paper feedback to the steering committee, included 3 keynote talks, 9 plenary presentations, 13 lightning talks, 4 panel discussions, and 6 breakout sessions. The hallmark of the meeting was the continuous active discussions from the entire audience, combined with note-taking by scribes working with the steering committee and public access to several shared documents so that all the participants could augment their comments in person with additional written notes. Strong engagement from both junior and senior participants was highlighted by the opening keynote given by Amir Gholami, UC Berkeley postdoc expert in machine learning, followed by two more keynote presentations by Flemming Crim, Chief Operating Officer for the National Science Foundation, and Henry Kautz, Division Director for Information & Intelligent Systems (IIS) at the National Science Foundation. Overall, the workshop provided a unique forum for interaction and frank discussions among AI experts, CI professionals, and domain scientists and engineers, which was appreciated by all participants.

Post-Workshop Activities / Deliverables

Workshop Report: The steering committee worked with the supporting editors and other community members to produce this report, which captures all the activities that led to the preparation of the workshop as well as the outcome of the numerous discussions. In particular, the overall community feedback is compiled in six main sections corresponding to the six final working groups. The executive summary also includes a summary of the overall findings and recommendations. This report will be released to the public and posted on the workshop website.

2. Cyberinfrastructure Requirements for AI-Based Scientific Applications

Goal: The goal of the working group was to explore the requirements for cyberinfrastructure to facilitate the application of AI in scientific applications, including hardware and software infrastructures, and data repositories to enable exploration of scientific and engineering on (future) NSF infrastructure.

Background: Fast development of AI has inspired a wide adoption of AI-related techniques in science and engineering. AI-ML-based techniques have been applied in many fields of science and engineering, including remote sensing, cosmology, energy, cancer research, machine design and control, etc. Great national and international enthusiasm for launching major AI-centric new R&D programs is evident, as multiple workshops or mini-symposia on the early adoption of AI techniques have been sponsored by various federal agencies within the U.S. and across the world. CI has become an indispensable tool for modern scientific discovery, and a smarter future CI is perceived as an accelerator or gap-bridger for solving the pressing challenges in science and engineering.

A smart CI can drive scientific discovery in two important ways. When AI is directly incorporated into the process of solving scientific problems, a smart CI needs to handle a variety of instruments, multiple interrelated data types and associated metadata, data representations, and workflows. Two exemplar scientific studies are showcased in the next paragraph to demonstrate the application of AI in scientific studies and the associated demands and problems. When scientific studies, both computational and experimental, are constrained by their effective deployment of the fast-evolving CI, a smart CI can bridge the knowledge gap between the CI experts and the domain scientists and accelerate the full cycle of scientific discovery through expedient harnessing of the most current computational power. This second aspect will be further discussed in the next section. Each scientific domain needs to identify the best strategy to leverage the rapid development of AI and smart CI, customized to their methodology and computational/experimental demands. At this early stage, exchanges of user examples and visions, as is happening in community workshops, have proven to be extremely inspiring for the entire scientific community to encourage adoption and customized development.

Figure 1 AI/Deep learning accelerates progress in predicting dangerous disruptive events in magnetically confined fusion plasmas with unprecedented accuracy and speed on modern supercomputers [1].

A clean and sustainable energy future remains one of the major challenges for modern society. Breakthroughs are urgently needed for traditionally challenged research fields such as fusion and clean combustion. Figure 1 shows a simulated plasma flow using Princeton's Fusion Recurrent Neural Network (FRNN) code. Fusion energy, once successfully implemented and controlled, could significantly reduce the stress of future energy demands. Spatial and temporal information is integrated using convolutional and recurrent neural network components for predicting disruptions in tokamak plasmas with unprecedented accuracy and speed on leadership-class supercomputers [1]. Physics-based interpretability is introduced to output not only when a "disruption score" for the probability of a disruption event but also when a "sensitivity score" in real-time to indicate the underlying physical reasons for the imminent disruption, which provides targeted guidance for control actuators upon implementation into a modern plasma control system. The convergence of advances in CI and AI is demonstrated by combining realistic pre-disruption classifier simulations on the Summit supercomputer using the exascale class GTC code with the AI workflow. With the continuous development of AI in computational studies of fusion, efficient and realistic control strategies based on advanced AI predictors are expected to emerge for optimization of performance and avoidance of disruptions in initial operations of ITER. Continuous vetting of stable, scalable, portable, control systems and associated methodology on existing tokamaks is also anticipated with the fast prediction offered by the AI-augmented simulations.

In addition to being incorporated into modeling and inference processes, AI/ML are also employed in other aspects of scientific studies, such as formulating efficient and accurate numerical methods. In many physical systems, the governing equations are known with high confidence, but direct numerical solution is prohibitively expensive. Turbulence, a phenomenon related to but distinct from chaos and with strong roots in engineering, is such a field. Direct solutions of the governing equations involve discretizing the computational domain into sub-domains down to the smallest scale in the dynamic system, which remains computationally prohibitive when the scale separation is large. This situation can be alleviated by deriving effective models to approximate dynamics below the grid-scale, such that the requirement for resolution is less stringent. The derivation of subgrid models is often impossible to perform analytically and requires deep physical understanding. A recent study showcased methods of data-driven discretization to attain the desirable accuracy where necessary. The method uses machine learning to systematically derive discretizations for continuous physical systems [1]. On a series of model problems, data-driven discretization gives accurate solutions with a dramatic drop in required resolution and, hence, the computational cost in solving the equations. Two representative challenges across many scientific domains are identified: speed constraints imposed by effective training of the neural network and the scaling constraints to go beyond one-dimensional toy problems. Solutions to address these challenges from the perspectives of both CI and AI are urgently needed.

The above examples are just two of the many recent developments in accelerating sustainable energy and complex multiscale physics; application of AI to numerous other scientific fields is ubiquitous, including cosmology, drug discovery and repurposing, cancer detection, additive manufacturing, and robotics. Although these studies represent fundamentally different science, their demands regarding cyberinfrastructure are often common: for example, CI with the capability to handle multiphysical simulations coupled with AI techniques that are interpretable by humans; scalable algorithms for massively parallel systems; and trained experts who understand AI, science, and CI simultaneously. Other bottlenecks in fulfilling the promise of AI as a science driver that are not discussed above include the lack of common standards and benchmarks in respective scientific and engineering fields, due to the early stage of adoption, difficulty of defining benchmarks, and proprietary protection. In these cases, exemplar NSF programs and projects that can be used as a model for addressing these challenges and technical bottlenecks would prove to be tremendously beneficial to the whole scientific community.

Based on the discussions during the workshop and feedback before and after from the larger scientific community, a list of findings is briefly described below, followed by recommendations.

Findings:

- **Support for the entire scientific discovery lifecycle is essential.**
  The applications of AI in science and engineering mostly focus on traditional model inference. However, the application of AI and smart CI to other largely manual, labor-intensive, and rate-limiting aspects of the decision-making processes is limited, e.g., in formulating questions; designing studies; organizing, curating, connecting, correlating, and integrating cross-domain data; and interpreting results, etc. Efforts in developing hypotheses with unbiased data, connecting AI tools with multiscale modeling, human-in-the-loop intelligence, mediated teams, and institutional collaborations, are also lacking.
- **Collaboration via combined teaming of domain, CS/AI, and CI experts is necessary.**
  Without sufficient support for ready-to-deploy CI tools/middleware, domain scientists often shoulder the burden of tool development and optimization by themselves in an inefficient way, which adds to the difficulty of human capacity development and hinders the progress of scientific exploration. On the other hand, long-term support for training students/postdocs/research scientists who have the expertise of domain sciences, software engineering, AI, and cyberinfrastructure is rare.
- **There is little support for scalable AI algorithms deployed in production.**
  Without the development of scalable AI algorithms, it is impossible to efficiently use existing and upcoming supercomputing AI infrastructure. Methods such as stochastic gradient descent and other related first-order methods that use only gradient information are popular, but they have numerous well-known drawbacks. In particular, these methods often require extensive hyperparameter tuning, and they often result in sub-optimal AI accuracy when trained at scale, in particular for problems such as those that arise in smart cyberinfrastructure. Current support for such an endeavor is inadequate.

- **Domain-specific training opportunities that include CS/AI features are lacking for students in science and engineering**.
  The need is clear for courses in the foundations of AI/CS to be incorporated in science/engineering disciplines. Such courses are offered for motivated students and users with the appropriate background and have proven to be successful and popular

Recommendations: The workshop identified five priority research directions necessary to empower NSF scientists.

- **Enable access to the NSF-funded cyberinfrastructure as an "experimental instrument" with freely available data**.
  Performance and behavior data monitored from applications, systems, and networks are valuable assets generated by NSF-funded cyberinfrastructure. Full visibility and instrumentation should be enabled to allow access to such data to facilitate development of AI for CI. Such efforts can be augmented by establishing platforms for data sharing with the broader CI research community.
- **Encourage and facilitate the application of AI to the full lifecycle of scientific discovery, to go beyond applying AI only to the model inference process in scientific discovery.**
  Support is needed for nimble and trustworthy data cyberinfrastructures that manage a variety of instruments, multiple interrelated data types and associated metadata, and data representations, processes, protocols, and workflows. The software and expertise that can address the needs of multiscale modeling and its coupling with AI across levels need to be developed.
- **Develop special funding mechanisms to encourage interdisciplinary teaming/collaboration.**
  Encouragement of interdisciplinary collaborative efforts through synergistic awards (e.g., AI Institutes or Centers of Excellence) will facilitate communication and collaboration across domains and provide direct support to domain scientists while connecting AI researchers and CI professionals to their specific needs. Funding support to develop AI middleware can be another solution to fill the gap between domain and CI scientists. The expertise of social scientists in the area of communication between different communities can be positively leveraged.
- **Support scalable AI software that goes beyond PyTorch and TensorFlow as part of an NSF smart cyberinfrastructure agenda.**
  Finding the right ML model and training it for a new task requires considerable expertise and extensive computational resources. Moreover, the process often includes ad hoc rules that do not generalize to different application domains. These problems have limited the applicability and usefulness of ML in science, especially for new learning tasks. Research investments in developing scalable solutions, such as second-order methods (e.g., subsampled Newton methods, stochastic Lanczos methods, matrix-free iterative methods, and subsampled trust-region optimization), are recommended. These methods have been

shown to perform competitively for a range of AI training and inference tasks, and they are particularly well suited to upcoming supercomputing hardware technologies.

- **Augment the science and engineering curriculum with NSF-supported short courses and MOOCs.**
  Understanding the latest developments in the theory and application of AI should be customized into the curriculum of respective scientific/engineering domains. In the short term, short summer courses and MOOC can be leveraged to train the workforce. Student fellowships such as the Frontera Computational Fellowship should be established to encourage the training of AI-driven scientific discovery. Deeper understanding from domain scientists is necessary in the longer term; the addition and reformation of existing mathematical or computational courses could be targeted.

3. Development of Next-Generation System Cyberinfrastructure Tightly Integrated with AI Technologies

Goal: The goal of this working group was to explore how AI, ML, and smart technologies can inform the design and operation of NSF cyberinfrastructure, both the hardware and software, in centralized facilities and at the edge.

This discussion was facilitated throughout the workshop by a panel, several invited talks, and two breakout sessions. The first breakout session focused primarily on the physical aspects of the "central" CI, inside the data center, with the second breakout focused more on "edge" issues, data acquisition, and software layers.

Background: NSF cyberinfrastructure is critical to the nation's research leadership in science and engineering. The NSF cyberinfrastructure portfolio encompasses a spectrum of powerful platforms, including Frontera and Stampede-2 (TACC), Bridges/Bridges-AI, and Bridges-2 (PSC), Comet and Expanse (SDSC), Ookami (Stony Brook University), and Jetstream (Indiana University). These platforms provide invaluable computational and data management infrastructure to the national research community of faculty, postdocs, graduate and undergraduate students, and even K-12 students. They also support US industry, within well-defined limits, by providing unique resources to increase their competitiveness, and extensive training programs supported by the NSF XSEDE program and using NSF CI platforms for exercises that uniquely drive workforce development. This is an extraordinary foundation, and the service providers hosting the NSF CI platforms are experts in system design, deployment, and operation. The NSF CI ecosystem provides exceptional value and scientific impact.

However, we now have the opportunity to do even better. This opportunity is driven by three factors: the greatly increased complexity of centralized cyberinfrastructure, a much greater diversity of user applications and their requirements for the underlying system, and the increasing importance of edge computing. These factors are addressed, in turn, below.

Centralized CI has increased greatly in complexity. Unlike the monolithic supercomputers of past decades, contemporary NSF-centralized CI resources are often heterogeneous, offering within the same platform multiple varieties of processors, accelerators, and memory configurations that are individually optimal for different applications and workflow components. Their heterogeneity is appropriate for supporting a wide range of applications (as described below); however, it presents challenges in efficient scheduling, measuring efficient use of system resources, improving energy efficiency, and optimizing software.

The diversity of applications has also increased greatly. NSF CI has grown to support numerous research communities, including many "nontraditional" communities that only recently have begun to use high-performance computing. For example, PSC's Bridges system has supported researchers representing 121 principal fields of study. That diverse user community now runs a

broad array of applications with different characteristics and requirements. Highly parallel, compute-intensive, physics-based simulation and modeling continues to be important, but is now only one aspect of the applications ecosystem. Equally important is applying AI to accelerate simulation and modeling by replacing computationally expensive phases with inferencing that can be many orders of magnitude faster with no loss of accuracy ("surrogate models"), AI applications requiring great amounts of resources for training and inferencing, scalable AI and ML for high-performance data analytics (HPDA), applications written in Python and other high-productivity languages, image and text processing, and containerized applications of many different kinds. This introduces challenges in optimally mapping applications onto resources, guiding users to the best resources for their applications, and helping them estimate the levels of resources they will require.

Edge computing adds a dimension to the complexities described above. Like centralized CI, edge computing encompasses a wide range of devices included in a single platform. Edge computing also must support a range of applications, and these applications add critical optimization criteria such as power consumption, memory footprint, and communications. Thus, the challenges that apply to centralized CI and diverse workloads apply also to edge computing and potentially without the substantial power of a centralized resource to leverage for solutions.

Clearly, traditional expertise in system operations and user support does not scale to the complexities of the heterogeneous centralized cyberinfrastructure, diverse application workload, and emerging edge computing of today and tomorrow. AI has great potential to apply cyberinfrastructure more efficiently and to better support users, which will translate to greater scientific impact. The field of AIOps [11], also known as SysML, has been growing rapidly and may provide great benefit for NSF cyberinfrastructure.

NSF CI service providers already collect a wealth of system instrumentation data from centralized CI. However, these data have been largely untapped, for a variety of reasons. Generally, the data have been collected for specific purposes within each site, so they reside in a variety of databases and formats, each well suited to its purpose, but needing substantial wrangling to be made interoperable. For example, individual sites have millions of job records, extensive system telemetry, and various performance data, each maintained separately. Fully preparing that data for analysis, even within one site, is human time-intensive. Furthermore, the metadata for system data have not yet been sufficiently standardized to bring data together from multiple service providers. However, this working group has recognized the great value in such standardization, so that AI techniques could be applied to improve system operations and the efficient use of system resources ("AIOps"), provide users with much-needed guidance regarding appropriate resources and resource estimation, and improve the mapping of applications onto resources. With sufficient data, we will also have opportunities to apply AI to design future cyberinfrastructure that will drive a new era of computational and data-driven science.

The rapid evolution of computer hardware, especially to enable specific fields such as AI, creates the need for the science and engineering community to gain experience with new technologies. Testbeds consisting of relatively modest resources and experimental platforms consisting of resources that make sense only at a larger scale have proven valuable in other contexts.

In summary, both NSF cyberinfrastructure and the application workload that runs on it have increased greatly in complexity. Service providers have considerable data on system activity, but without an appropriate, shared metadata framework. Nonetheless, AIOps has great potential to improve system efficiencies and to guide users if the data can be properly curated. Testbeds and experimental platforms can play a vital role through broadening community experience with new technologies and informing the design of future systems, including, potentially, through AI techniques.

Findings: The working group identified five broad findings relating to AI and smart CI technologies:

- **Numerous untapped opportunities exist for AI technologies to improve the NSF cyberinfrastructure.** AI can address a number of challenges from the system infrastructure perspective of a smart CI, such as enhancing reliability, security, resilience, performance, energy efficiency, energy utilization, and usability. Cyberinfrastructure providers have, and can potentially provide, extensive system data, ranging from system health to performance counters. The data for each system must be aggregated into multiple databases, and the first stage of curating such data is typically manual. User privacy must be respected, requiring the data to be appropriately deidentified.

- **Architectural changes have enormous promise to reduce the time required for AI computations.** With the increasing computational demands of AI algorithms, training time and energy efficiency are both extremely important. Emerging architectures that are carefully tuned to deep learning training potentially offer orders of magnitude improvement. The potential of new architectures for science and engineering requires benchmarks from those developed primarily for social network and business uses. Testbeds and experimental platforms can range from modest through substantial investments, the latter for technologies where the unit cost and potential value are both substantial.

- **AI can help users make good decisions about how much power their computations cost.** Maintaining awareness of overall energy utilization (beyond energy efficiency) is also important. The top priority of scientific impact should be pursued without losing sight of its carbon footprint. AI can help lower energy utilization, in some cases by many orders of magnitude, by replacing computationally intense simulations with surrogate models, reduced-order models, or other shortcuts. Furthermore, new architectures can reduce energy utilization by additional orders of magnitude.

- **The lack of skilled individuals who can bridge from infrastructure to application, especially in CI, is a factor limiting US progress, in industry and academia.** Noted primarily by the cloud providers in the group, human capacity development and retention are critical. We need to build communities and interdisciplinary expertise that bridges domains and AI. Although an acute problem in the academic research enterprise, this problem is perhaps more acute in the industry — salaries are not the issue; we do not have enough individuals to go around, at virtually any price.

- **At the edge of our CI are numerous pain points concerning how data can be acquired to drive models and take actions.** Concerns range from security and privacy to reliability and sharing. Standards for sensors are lacking at the edge — from ubiquitous tasks like authentication and validation to more complex areas like sharing data streams with multiple applications. Control and management of sensors are complicated, often making the assumption that relatively frequent manual intervention is acceptable for calibration, maintenance, resets, etc. Sensors are unreliable and often inaccurate, which without standards is difficult to quantify.

Recommendations: The workshop identified five priorities to drive scientific and engineering discovery through smart CI:

- **Initiate a coordinated program through which CI providers make available curated system data.** Data including job execution, I/O patterns, performance counters, and other quantities can be made available through modest curation, development of an appropriate controlled vocabulary, documentation, and potentially training and tools to facilitate access. The resulting datasets will then be of great value to optimize existing systems, design new ones, improve usability, and, eventually, develop self-optimizing systems.
- **Provide mechanisms for testbeds and experimental platforms to evaluate their effectiveness for science and engineering and to build community experience.** "Testbeds" of small to moderate scope should be frequently refreshed. "Experimental platforms" should potentially be larger and longer term, to access technologies that have higher unit costs together with great potential. Each platform could possibly leverage public/private partnerships. Testbeds and experimental platforms should address important nontraditional application areas such as deep learning and graph analytics; energy-aware computing for evaluation of energy-efficient approaches, allowing exploration of constraining or guiding energy utilization; and smart sensors and other devices for edge computing, allowing the development of methods for AI-ready sensors and edge processors. Repositories of best practices between academia and industry should be created. "AI-defined CI experiments" should be supported, where a smart CI goal can be specified and transformed into a finite set of experiments to be evaluated and tested.
- **Broaden education and training opportunities to develop domain/AI interdisciplinary expertise.** Interdisciplinary research should be fostered to expose AI students to domain science and vice versa.

4. AI/ML Improvements on NSF CI

A number of example resources and scientific applications have successfully deployed artificial intelligence and machine learning (AI/ML) techniques as a part of the solution. However, these solutions are custom-built for the problems. We have several challenges in scaling and generalizing the developed methods and tools at the CI level. Thus, the goal of this working group was to discuss the state of the art in this area and explore potential improvements needed in AI/ML algorithms, software infrastructure, and data repositories to enable effective use of these technologies on (future) NSF infrastructure.

Section 2 of this report outlined examples and needs for using AI and ML in scientific applications at scale. The summarized discussion demonstrated the impact of efforts not only on the application of AI/ML with domain expertise but also on the scalability of the developed methods for existing CI. Research teams still have difficulty deploying and using AI and ML in their scientific and engineering process without such significant effort.

The panel discussion at the workshop discussed the state of the art in AI/ML within the existing CI ecosystem and technologies and explored potential improvements needed in AI/ML algorithms, software infrastructure, and data repositories to enable effective use of these technologies for (future) NSF infrastructure. This panel discussion captured some technology disruptors related to the deployment of hardware and software.

The panel was asked about the state of the art in support of AI, ML, and smart computing as well as examples of any CI for AI within external and industry communities that could set an example for NSF. The panel also explored the state of the art, pain points, and opportunities for the deployment of tools, frameworks, and algorithms at scale and in production pertaining to the usability (both interactively and programmatically) by NSF's user communities. Lastly, the panel was asked for their recommendations on how to create an NSF-lead team science environment driven by the needs of AI-integrated applications and teams collaborating to design such applications.

An important part of the panel and the related breakout discussion focused on the importance of data flows and related workflow tools and software/systems engineering (DevOps) that can support AI/ML as a part of the bigger picture, including the coupling of systems, on-the-fly training, outer loop optimization, and large-scale deployment and utilization of AI infrastructure. Another important discussion point was related to how an NSF-scale AI infrastructure would be different compared to the industry-driven cloud-based AI platforms. It is important to focus on the needs of the science community that are not commonly supported in commercial applications and determine how to achieve economy of scale while using the industry resources.

The discussion concluded by noting the open challenges and primary cyberinfrastructure research areas summarized in the following findings.

Findings:

- **Applications need better support to integrate cutting-edge hardware and software.** Even where applicable technology exists, the effort needed to integrate these advances with the already complex scientific applications often proves insurmountable, which significantly hampers progress and wastes resources.
- **Current middleware and computer science infrastructure are not optimized for data-centric workloads.** A number of system-related parts of the software stack, such as data management, communication, file IO, etc., have been heavily optimized for simulation-driven workflows and often do not support the more complex, data-driven needs of AI.
- **The complex iterative workflows necessary for AI do not map to existing batch-focused resource management.** Applications increasingly consist of complex workflows incorporating heterogeneous CI capabilities, i.e. iterative or recursive scheduling, drastically varying job size, an unprecedented number of jobs, etc., that are not compatible with existing batch scheduling, resource allocations, security postures, etc.
- **Existing toolchains are made by and for ML experts.** The currently available tools, although powerful, are driven almost exclusively by commercial applications and are used by ML experts. Application communities have faced challenges to build up sufficient expertise to deploy these solutions and, once built, to correctly and confidently interpret the results.

To address the four high-level findings outlined above, the workgroup participants recommend four corresponding directions for improvement.

Recommendations:

- **Low-Level Interfaces:** One of the challenges in applying AI technology is the incredible pace of innovation in both hardware and low-level software frameworks. Specialized hardware such as GPUs, TPUs, and other chip designs (e.g., Cerebras: cerebras.net/) as well as high-performance software frameworks, e.g. Horovod (github.com/horovod) and Aluminum [12], have the potential to accelerate the AI workflow. However, continuously integrating new technology into existing applications places an unsustainable burden on developers. As a result, domain scientists either spend a significant portion of their resources addressing low-level software challenges or forgo the potential performance gains by using outdated technology. Instead, we recommend research focus on developing transparent abstractions between low-level system libraries and specialized hardware to simplify sustainable software development. In particular, the focus should be on flexible, portable, and scalable solutions that allow applications to seamlessly transition from local computing resources to leadership facilities and cloud computing while maintaining high performance.
- **AI-Aware Middleware:** Efficiently applying AI algorithms at scale depends on a variety of general-purpose capabilities, such as flexible workflows, scalable data stores, or efficient IO infrastructure. However, in many cases, AI/ML algorithms result in different usage patterns and bottlenecks compared to traditional simulation codes. For example, many AI workflows consist of a complex interactive assembly of different tools that must

be coordinated and connected, potentially across different facilities, i.e. to integrate streaming experimental data. Such solutions are often not compatible with existing batch scheduling, resource allocations, or security postures. Similarly, the random accesses inherent in stochastic gradient descent algorithms represent the worst-case behavior for file IO or data accesses in general. Instead, we recommend research focus on a new generation of AI/ML-aware middleware solution that adapts, extends, and optimizes existing frameworks for AI/ML algorithms as well as the corresponding hardware and software stacks.

- **Solution Environments:** Currently, the use of machine learning techniques requires significant programming skills and advanced knowledge of the underlying methodologies. For example, developing efficient data representations, choosing appropriate model architectures, or tuning various hyperparameters all demand a detailed understanding of machine learning fundamentals, unfamiliar to many application scientists. This knowledge gap results in a significant bottleneck in the widespread use of AI techniques. To address this challenge, we recommend a research focus on high-level solution environments and next-generation workflow tools that provide easily accessible and composable components for various levels of expertise. For example, we need a (community-vetted) repository of standard AI models for different use cases and data types for nonexperts to apply. At the same time, solutions must be flexible and extendable to accommodate future applications. Finally, the access should be tiered, enabling users ranging from AI professionals, to experienced practitioners and nonexpert users access according to their knowledge and abilities.
- **Responsible AI:** An often overlooked aspect of applying AI/ML technologies is the difficulties stemming from the complex, black-box nature of the models. Even experts find it challenging to confidently judge whether or not enough training data are available, what a model has learned, and how reliable any given model may be. This challenge is exacerbated by the lack of theoretical guarantees on the stability and generalizability of data-driven models. Consequently, interpreting results, understanding uncertainties inherent in a given solution, or recognizing potential biases in the data are all difficult if not impossible for application-focused researchers. To address this need, we recommend a research focus on both fundamental advances in machine learning to address issues such as interpretability, uncertainty quantification, privacy, etc. and the development of tools to make the research results accessible to nonexperts. Ultimately, the tools developed for this recommendation should be directly integrated into the solution environment discussed above to provide not only the means to produce results but also ways to understand the context in which these results were achieved, correctly interpret the findings, and check for biases and other potential pitfalls.

5. Improving AI-Associated Data Repositories on NSF-Enabled Cyberinfrastructure

The working group discussed the opportunities for building shared data and model repositories for training and inference, with applications in computer science, collaborative problem solving, and evidence-based CI development. Major advances in applications, infrastructure, and collaborative discoveries can be effective in the collection and use of data to further understand the effectiveness of learning algorithms, hardware, middleware, user interfaces, and teamwork. These advances includes repositories of scientific data generated on the NSF-supported CI as well as the large amounts of data collected while monitoring the scientific computations that can be used to enable AI control and optimization of the operations. In both cases, we aim at developing standards and repositories that facilitate data sharing and retargeting the information available for new AI tasks.

The panel and the associated breakout group were asked to discuss existing repositories of training data in support of AI within and outside NSF. Example programs and projects can be used as a model for creating a training data sharing ecosystem and the building blocks of such an ecosystem. Some of the discussion was related to human capacity development issues regarding (i) the ability of individuals and teams in the NSF user community to find, access, interpret, evaluate, and validate data and AI/ML capabilities and (ii) cultural issues regarding data and model sharing.

Findings

- **Convergence research requires middleware for integration and scalability of AI services on cyberinfrastructure, but not having integrated CI performance data repositories hinders such research.** Data-driven performance prediction and modeling work service and integrated applications are key to achieving smart CI capabilities at scale. Dynamic integration of system data with ML capabilities enables AI for smart middleware and workflow steering. However, the research of performance prediction currently depends on standalone databases collected and compiled by individual researchers and collaborators in supercomputer centers, making these data collections unavailable to the research community at large. Additionally, the collected performance and other related CI data lack standards, making the collected datasets often not interoperable. Open and community-driven archives of CI data are needed for increased research in smart CI middleware.
- **AI has the potential to add intelligence to CI and associated computer science research as well as other domain applications.** AI/ML-integrated science often involves a team of scientists with cross-disciplinary expertise collaborating and communicating to solve a problem. The process might involve an interactive exploration of historical and real-time data, access to active storage systems, and the networking and computing resources available to the team. A smart CI should also enable the betterment of the science of collaboration applied to application-driven problem-solving.
- **AI techniques requiring a large amount of data and labeled training datasets benefit from shared infrastructure for data and models.** Training data for AI/ML research are

an integral part of the at-scale applicability of AI/ML by diverse users. However, creating and maintaining high-quality labeled datasets can be costly and difficult to maintain as a service.

- **New CI that builds on the analysis of trends in utilization of CI systems and that provides transparent data on system and application performance in real-time can be prioritized.** Major advances in AI/ML-integrated science can result from building the right computing infrastructure for the evolving needs of applications. New infrastructure decisions and priorities require AI-based techniques and data from existing systems to come together. However, current data from CI systems and their utilization are kept in silos that are hard to access and analyze.
- **The current CI culture is not built around data sharing, but the data have the potential to enable new innovations and improve teaching when available.** The above-mentioned siloed culture and organization related to CI use and system performance hinders the data-driven analysis of CI and, in turn, the development of smart CI techniques and middleware. The added workload to manage and openly share infrastructure is not justified by the incentive structure. Most often cyberinfrastructure providers do not directly have a teaching-related mission, and various contributions to computer science training are not directly supported by their funding.

Recommendations

1. **NSF should support cross-cutting programs for the interoperability of CI data, methods, and models at scale.** Standards development for data related to CI is a potential enabler for interoperability and should be funded along with data infrastructure that ensures collection, curation, integration, and management of CI data in the long term. The data share should conform to FAIR data principles as well as enable programmatic access to data in real-time for intelligence and responsiveness of CI systems, services, and middleware.
2. **Research in support of interactions of collaborative groups with smart AI systems is required for the application of AI to problem-solving with scientific, societal, and educational impact.** There is a need to develop research formalisms toward translational tools and environments for teams to integrate individual work in a trustworthy and responsible way into the scientific problem-solving process. The transition from exploration to scalability requires data collection within the CI of the performance, accuracy, and scientific integrity of each step as well as the integrated solutions. Research on dynamic utilization of AI should be encouraged to close the gap between developing individual steps and integrating team activities within solutions.
3. **The community needs to develop middleware that can federate and integrate open, shared, well-annotated data and model repositories for training and validation, with the possibility of intradisciplinary and interdisciplinary knowledge, model, and method transfer.** Domain knowledge and annotations are needed for the curation of training data and models for smart CI. New standards for sharing and curating CI system data are needed. CI providers should participate in regular data curation sessions for the

development of open training datasets for smart CI development. Additionally, data repositories that provide access to historical and real-time usage data will enable smart CI middleware and tool development.

4. **New CI research methods and environments need to be developed for the systematic generation of digital twins for virtual experimental CI laboratories and instruments.** Monitoring and managing NSF CI systems and facilities require having access to data from different parts of the infrastructure, including physical and digital components. Projects to generate digital replicas should be funded for community-driven progress toward smart CI systems and environments as well as more efficient facilities. Balancing privacy, transparency, and openness of CI-related data is an important research problem that needs to be addressed.
5. **NSF should develop programs that facilitate, incentivize, and reward data sharing among CI system providers.** The abovementioned smart CI expectations add new complexities to the operational workflow of CI and facility providers. An updated funding model and/or incentive structure needs to be developed to catalyze and encourage the adoption of smart CI by system providers. At the minimum, smart CI should be added as a funded component of existing systems. However, connecting it as a function across systems is likely to have a bigger impact on new innovations.

6. Human Capacity Development in AI and Smart Cyberinfrastructure

Goals. The goal of this working group was to identify the existing challenges and proposed solutions associated with the creation of a sustainable and persistent generation of experts in AI, ML, big data, and CI. Dedicated discussions took place in a panel and a breakout session; eventually, the discussion was facilitated throughout the workshop discussions across the other workshop components.

Background. The workshop pointed out how, across an existing rich body of work, there is the tendency to use different terms and languages to discuss similar efforts targeting the creation of a solid, persistent generation of experts in CI and how most efforts, while successful, are compartmental in the sense that they tackle one aspect of a more complex problem (e.g., only the curriculum development at the undergraduate or graduate level). The workshop started with the search and definition of a general term under which to include all aspects of human-capacity building, including the education, training, and forming of transdisciplinary, diverse experts who synthesize previously disjointed fields of science and engineering, and converged toward the term *human-capacity development.*

Although specifically addressed by a working group at the workshop with a panel and a breakout session, the need for improved human-capacity development for smart cyberinfrastructure (CI) was a transversal topic discussed across all components of the workshop. The workshop concurred on the fact this is a pressing problem for science, education, and industry alike. The workshop participants also concurred that this problem is multidimensional with a variety of significant stakeholders, including:

- the diverse individuals participating in the process, including students, staff, and research professionals at every level of expertise;
- a wide spectrum of science and engineering communities needing people with these skills, and
- the academic, governmental, and industry sectors in which these research activities are embedded.

To sum up, workshop participants identified the urgency for (i) the creation and maintenance of capable, diverse, and multidisciplinary human capacity (i.e., including students, staff, and research professionals at every level of expertise) and (ii) the creation of laboratory organizations to facilitate the substantial transformation of the values that inform the profession (i.e., from the number of papers written and students graduated to the number and type of intellectual property generated, the amount of software released and data acquired, and the number skilled professionals

produced). The participants concur on the fact that the success of a capable human capacity for smart CI can be achieved only with a broad coalition of stakeholders (including universities, industry, current cyberinfrastructure providers, multiple government agencies, and the international community).

Findings: The working group identified two broad findings relating to workforce capability development:

- **Challenges faced in human capacity development.** Experience shows that human-capacity development confronts a number of critical and diverse problems when dealing with creating and retaining capable individuals. These problems are fivefold. First, the pace of both intellectual and technological change in the area of smart CI outstrips the ability of academic institutions to adapt their curricula. Second, continuous retraining is required because of the rate of technological change, but this retraining does not fit in the traditional education model. Third, using smart CI for sciences and engineering is inherently a multidisciplinary, cooperative activity that requires communication and teamwork. Fourth, the different groups and individuals have different entry-level thresholds; the learning requirements are often domain-specific. Last, training-by-doing and on-the-job-learning are generally required.

- **Transformation of values that inform the profession.** We have observed a substantial transformation of the values that inform the profession. Such a transformation, however, has a different pace for the emerging CI community in contrast to the traditional academic community. The productivity metrics in the CI community have moved from the number of papers written and students graduated to the number and type of intellectual properties generated, the amount of software released and data acquired, and the number of skilled professionals produced. The traditional academic business models, however, have not evolved at the same pace and do not support and, in fact, are often in conflict with this transformation of values. Addressing this dichotomy is a problem for the international community, including academia, professional societies, national labs, and government agencies.

Recommendations: The workshop identified two priority recommendations necessary to build an adequate pipeline of capable scientists and engineers for smart CI.

- **Establish a capable, diverse, multidisciplinary human capacity.** A broad coalition of stakeholders (including universities, industry, current cyberinfrastructure providers, and multiple government agencies) needs to address the creation and maintenance of a capable, diverse, multidisciplinary human capacity:
    a. To prevent obsolescence due to the pace of technology changes – universities

should refocus the core curricula around the foundational areas of mathematics and statistics (e.g., linear and tensor algebra, graph theory, statistics) that underlie all the different flavors of AI (e.g., neural network, machine learning, data mining) and are technology agnostic.

b. To support multidisciplinary needs — universities should adapt traditional subject matters for a transdisciplinary audience (e.g., efforts at the University of Chicago to develop a computer architecture course for nonengineering students).
c. To support different entry levels — smart CI facilities should focus on sustainable, common, cross-community interfaces for deployable scalability (for example, by leveraging the Jupyter Notebook or derivative tools such as Google's Collaboratory, Microsoft's Azure Notebook, and AWS' SageMaker Notebook).

- **Create new laboratory organizations.** The same broad coalition of stakeholders needs to address the creation of laboratory organizations to facilitate the transition to the new productivity model and prevent rollback to old ones. The creation process must consider these key aspects:
    a. The organizational design must be addressed from all the institutional loci that feed into that smart CI community, such as academia, professional societies, national labs, and government agencies, and across national boundaries.
    b. NSF should help stakeholders create new models of laboratory organizations with corresponding acknowledgment and support for individuals beyond / other than the postdoc position (e.g., research directors). The new organizational structures should reflect the metrics of discovery, validation, collaboration, and human-capacity building.
    c. NSF should help stakeholders develop a reward system to facilitate and sustain the transition.
    d. To oversee the implementation of these recommendations and modify them as technologies continue to rapidly evolve, NSF should appoint a jointly managed federal, higher education, and industry commission. They should report at least annually to the president's science advisor.

7. International Collaborations

Goal. The working group targeting human-capacity development also discussed existing challenges and proposed solutions associated with the leverage of international collaboration for the development, use, and sharing of AI, ML, and smart technologies. This discussion was facilitated throughout the same panel and breakout session hosting the discussion on human-capacity development.

Background. International collaborations promoting smart CI interoperability and sharing across international groups are a vital part of successful domestic research built on AI, machine learning, and big data. To this end, we need a general framework for international investment in global science and engineering that enables international researchers to identify potential collaborators and create trust among international collaborators through multiphase project organizations.

The international nature of scientific collaborations requires smart CI interoperability and sharing across groups and countries; successful international collaborations add value for all parties in research and productivity activities (e.g., people, data, and software and hardware). The international community faces several challenges in the pursuit of such collaborations. In some cases, the already difficult access and use of smart CI services are exacerbated by factors such as: (i) the services are provided by industry subject to governments' access limitations and restrictions, and (ii) the facilities are located in government buildings to which international collaborators have limited access. In other cases, services to support interoperability and sharing are not provided at all because, for example, they are not seen as being profitable. The ability of the smart CI community to excel depends on our ability to build upon existing international best efforts.

Findings. The working group identified two key findings relating to existing international efforts:

- **Existing successful efforts.** Existing international initiatives are already serving as successful models for sustainable collaborations. Examples of these initiatives include PRAGMA, CENTRA, and BDEC/IESP, among others. Moreover, NSF has successful program models through its office of International Science and Engineering, including PIRE, US-Japan, and US-Israel.
- **Existing successful NSF initiatives.** Current NSF initiatives for project organizations are also promising solutions to serve the needs of an international platform. These initiatives include the current augmentation of NSF annual reports to make the impact and outcome of projects searchable (e.g., through the collection of projects' metadata), and the introduction of two-phase projects (i.e., a one-year planning project followed by a possible multiyear research project) to facilitate team building.

Recommendations. The international efforts discussed above should be further extended by pursuing the following recommendations:

- **Provide a framework for international investment.** Such a framework should be co-located in a global science and engineering AI, machine learning, big data, and smart CI ecosystem, for example, a National Academies-managed joint higher education and industry consortium to (i) develop new visions, (ii) develop strawman roadmaps, (iii) define suitable collaborative mechanisms (e.g., MOU, by-law), and (iv) define sustainable business models (e.g., persistent funding).
- **Enable international researchers to identify potential collaborators.** This objective can be achieved by making different national research projects searchable and easy to access and by leveraging improvements in NSF annual reports, extending the current reports with richer searchable metadata.
- **Establish a multiphase project organization.** Such an organization can thrive if supported by coordinated solicitation and awards at the international level. The organization should build mechanisms aimed at creating trust among international collaborators, to allow coordinated project progress and to enable sustainable funding.

8. Acknowledgments

We would like to acknowledge the contributions of Brenda Peterson of the Scientific Computing and Imaging Institute (SCI) for her help with the workshop organization. We would like to acknowledge the contributions of numerous note takers, who, during the meeting, made it possible to keep track of the conversations. We would like to thank Christine Pickett from the University of Utah for editorial help with the report. Finally, we would like to thank Nathan Galli for putting together the workshop website. The workshop was supported by the National Science Foundation through grant number OAC 1941085.

9. References

[1] J. Kates-Harbeck, A. Svyatkovskiy, and W. Tang, "Predicting disruptive instabilities in controlled fusion plasmas through deep learning," Nature, vol. 568, no. 7753, pp. 526–531, Apr. 2019.

[2] S. Shadley, "Flash Memory Summit 2019: The Year of Computational Storage." 2019 [Online]. Available: https://www.ngdsystems.com/page/Flash-Memory-Summit-2019:-The-Year-of-Computational-Storage

[3] T. Omura, "Data-Driven Applications Gravitate to Computational Storage." Scale Flux, 2019.

[4] K. Hao, "Training a single AI model can emit as much carbon as five cars in their lifetimes," MITS Technol. Rev., 2019.

[5] Google, Ed., "Moving towards 24x7 carbon-free energy at Google data centers: Progress and insights," Oct. 2018.

[6] F. Yang and A. A. Chien, "Large-Scale and Extreme-Scale Computing with Stranded Green Power: Opportunities and Costs," IEEE Trans. Parallel Distrib. Syst., vol. 29, no. 5, pp. 1103–1116, May 2018.

[7] Executive Order on Maintaining American Leadership in Artificial Intelligence, Issued on: February 11, 2019. https://www.whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/

[8] The COVID-19 High Performance Computing Consortium. https://covid19-hpc-consortium.org/

[9] Scientists worldwide apply supercomputing resources to save lives and accelerate search for cure. https://www.tacc.utexas.edu/covid-19.

[10] Jan Rowell, Supercomputers Take on COVID-19, https://medium.com/@janrowell/supercomputers-take-on-covid-19-96a27b64b02.

[11] G. Blokdyk, AIOps A Complete Guide, 5STARCooks, March 18, 2019.

[12] N. Dryden, N. Maruyama, T. Moon, T. Benson, A. Yoo, M. Snir, and B. Van Essen, Aluminum: An Asynchronous, GPU-Aware Communication Library Optimized for Large-Scale Training of Deep Neural Networks on HPC Systems, 2018 IEEE/ACM Machine Learning in HPC Environments (MLHPC), Dallas, TX, USA, 2018, pp. 1-13, doi: 10.1109/MLHPC.2018.8638639.

Appendix A

Workshop Participants and Contributors

	Name		Affiliation
1	Ilkay	Altintas	San Diego Supercomputer Center
2	Fatima	Anwar	University of Massachusetts
3	Mark	Asch	University of Picardy, CNRS
4	Aditya	Bhaskara	University of Utah
5	Karan	Bhatia	Google Cloud
6	Sanjukta	Bhowmick	University of North Texas
7	George	Biros	The University of Texas at Austin
8	Peer-Timo	Bremer	Lawrence Livermore National Laboratory
9	Paola	Buitrago	Pittsburgh Supercomputing Center, Carnegie Mellon University
10	Sunita	Chandrasekaran	National Science Foundation
11	Vipin	Chaudhary	National Science Foundation
12	Andrew	Chien	University of Chicago
13	Carlos	Costa	IBM Research
14	Fleming	Crim	National Science Foundation
15	Rupert	Croft	Carnegie Mellon University
16	Ewa	Deelman	University of Southern California Information Sciences Institute
17	Adrian	Del Maestro	The University of Vermont
18	Deborah	Dent	Jackson State University
19	Tiziana	Di Matteo	Carnegie Mellon University
20	Nicola	Ferrier	Northwestern University, ANL
21	Jose	Fortes	University of Florida
22	Ian	Foster	University of Chicago
23	Geoffrey	Fox	Indiana University

24	Amir	Gholaminejad	University of California, Berkeley
25	Helen	Gu	North Carolina State University
26	Salim	Hariri	The University of Arizona
27	Hank	Hoffmann	University of Chicago
28	Vasant	Honavar	Penn State University
29	Eliu	Huerta	NCSA/University of Illinois at Urbana-Champaign
30	Olexandr	Isayev	Carnegie Mellon University
31	Henry	Kautz	National Science Foundation
32	Kerk	Kee	Texas Tech University
33	Sidharth	Kumar	University of Alabama at Birmingham
34	Yong	Liu	New York University
35	Ling	Liu	Georgia Institute of Technology
36	Anirban	Mandal	Renaissance Computing Institute (RENCI), UNC Chapel Hill
37	William	Miller	National Science Foundation
38	Jelena	Mirkovic	USC Information Sciences Institute
39	Terry	Moore	University of Tennessee
40	Pania	Newell	University of Utah
41	Nicholas	Nystrom	Pittsburgh Supercomputing Center, Carnegie Mellon University
42	Manish	Parashar	National Science Foundation
43	Valerio	Pascucci	University of Utah
44	Abani	Patra	Tufts University
45	Steve	Petruzza	University of Utah
46	Maryam	Rahnemoonfar	University of Maryland
47	Glenn	Ricart	University of Utah and US Ignite
48	Morris	Riedel	Juelich Supercomputing Centre
49	Naveen	Sharma	Rochester Institute of Technology
50	Dan	Stanzione	The University of Texas at Austin

51	Brian	Summa	Tulane University
52	Yoshio	Tanaka	National Institute of Advanced Industrial Science and Technology (AIST)
53	Bill	Tang	Princeton University
54	Michela	Taufer	University of Tennessee
55	Pavan	Turaga	Arizona State University
56	Bhuvan	Urgaonkar	The Pennsylvania State University
57	Duminda	Wijesekera	George Mason University
58	Karen	Willcox	The University of Texas at Austin
59	Yanfang (Fanny)	Ye	Case Western Reserve University
60	Xinyu	Zhao	University of Connecticut

Appendix B

Workshop Program

2020 NSF Workshop on Smart Cyberinfrastructure

February 25-27, 2020, Hyatt Regency Crystal City, VA

Rooms: Plenary sessions in Conference Theater

Breakout 1 in Potomac I

Breakout 2 in Potomac VI

Breakout 3 in Potomac V

Day 1 – Tuesday, February 25, 2020

07:00 – 08:00 Registration/Breakfast

08:00 – 08:10 ***Valerio Pascucci,*** *University of Utah & Steering Committee*
Introductions, structure of the workshop, expected outcomes

08:10 – 08:40 ***William Miller,*** *Science Advisor, NSF Office of Advanced Cyberinfrastructure (OAC)*
Manish Parashar, *Director, NSF Office of Advanced Cyberinfrastructure (OAC)*
NSF Welcome and workshop introduction

08:40 – 10:00 Setting the stage: Steering Committee (10 min presentations + 5 min questions)

- ***Ilkay Altintas,*** *San Diego Supercomputer Center*
  Improving AI Algorithms and Associated Data Repositories on NSF-Enabled Cyberinfrastructure
- ***Dan Stanzione,*** *The University of Chicago*
  Improve Efficient Use of the NSF-Enabled Cyberinfrastructure Through AI/ML and Smart Technologies
- ***Xinyu Zhao,*** *University of Connecticut*
  Cyberinfrastructure Requirements for AI/ML-Based Scientific Applications
- ***Michela Taufer,*** *The University of Tennessee, Knoxville*
  Large Collaborative Initiatives and Engagement of the International Community
- Open discussion (20 min)

10:00 – 10:15 Coffee break

10:15 – 10:45 Keynote talk 1, ***Amir Gholami,*** *UC Berkeley*
An Integrated Approach for Efficient Neural Network Design, Training, and Inference

10:45 – 11:45 15 min plenaries

Plenary talk 1, ***Karan Bhatia,*** *Google*
The Google Cloud

Plenary talk 2, ***Yoshio Tanaka,*** *AI Bridging Cloud Infrastructure (ABCI)*
Challenges for Smart Cyberinfrastructure in Making Society 5.0 a Reality

Plenary talk 3, ***Bill Tang***, *Princeton Plasma Physics Laboratory*

Features of an NSF Smart Cyberinfrastructure Roadmap Supporting Science Applications

Plenary talk 4, ***Mark Asch****, University of Picardy*
Smart CI and the Real World

11:45 – 12:00 Keynote talk 2, ***Fleming Crim,*** *Chief Operating Officer, National Science Foundation,*
The NSF Vision

12:00 – 13:00 Lunch break

13:00 – 14:00 Panel discussion 1: AI/ML Opportunities for a Modernization of the NSF Cyberinfrastructure from the Edge to Large Facilities
Organizer: **Dan Stanzione**
Panelists: **Ian Foster, Salim Hariri, Jelena Mirkovic, Helen Gu**

14:00 – 14:15 Organize breakout sessions

14:15 – 15:15 Breakout sessions start parallel discussions:
Breakout 1 in Potomac I: Cyberinfrastructure Requirements for AI/ML-Based Scientific Applications (Lead **Xinyu Zhao,** co-leads/scribes **Jose Fortes, Helen Gu**)
Breakout 2 in Potomac VI: Development of the Next Generation Physical (Hardware) Cyberinfrastructure Tightly Integrated with AI, ML and Smart Technologies (Lead **Dan Stanzione**, co-leads/scribes **Ian Foster, Steve Petruzza**).
Breakout 3 in Potomac V: Improving AI/ML Algorithms on NSF-Enabled Cyberinfrastructure: (Lead: **Ilkay Altintas**, co-leads/scribes **Michela Taufer, Timo Bremer**)

15:15 – 15:30 Coffee break

15:30 – 16:00 Breakout sessions reconvene and complete answers to main charge questions

16:00 – 16:15 Breakout leads summarize findings and recommendations in a set of slides and a shared document

16:15 – 17:00 Leads report to the entire workshop (15 min each group including questions)

Day 2 – Wednesday, February 26, 2020

07:00 – 08:00 Breakfast

08:00 - 08:10 **Valerio Pascucci,** *University of Utah & Steering Committee*
Summary of Day 1 / Goals of the Day 2

08:10 – 08:30 Plenary talk 5, ***Nicola Ferrier***, *Argonne National Laboratory*
AI@Edge: a Software-Defined Sensor Network

08:30 – 09:40 Lightning talks session 1

10 min each (5 slides + panel)

09:40 – 10:00 Plenary talk 6, ***Hank Hoffmann****, The University of Chicago,*
AI/ML Methods for Computer Systems Optimization

10:00 – 10:15 Coffee break

10:15 – 11:15 Lightning talks session 2

10 min each (5 slides + panel)

11:15 – 12:15 Panel discussion 3: Improving AI/ML Algorithms and Associated Data Repositories on NSF-Enabled Cyberinfrastructure
Organizer: **Ilkay Altintas,**
Speakers: **Peer-Timo Bremer, Yoshio Tanaka, Andrew Chien, Karan Bhatia**

12:15 – 13:10 Lunch break

13:10 – 14:10 Panel discussion 4: Effective Workforce Training, Development of a Community of Researchers and Practitioners
Organizer: **Michela Taufer**
Panelists: **Abani Patra, Jose Fortes, Mark Asch, Morris Riedel, Carlos Costa**

14:10 – 14:15 Organize breakout sessions

14:15 – 15:15 Breakout sessions start parallel discussions:

Breakout 4 in Potomac V: AI, ML and Smart Technologies for Data Acquisition, Experiments, and the Edge of the Cyberinfrastructure (Lead **Hank Hoffmann**, co-lead/scribe **Salim Hariri, Steve Petruzza**)
Breakout 5 in Potomac I: Improving AI/ML- Associated Data Repositories on NSF-Enabled Cyberinfrastructure (Lead **Ilkay Altintas**, co-lead/scribe **Helen Gu, Timo Bremer**)
Breakout 6 in Potomac VI: Effective Workforce Training, Development of an International Community of Researchers and Practitioners (Lead **Michela Taufer**, co-lead/scribe **Terry Moore, Jose Fortes**)

15:15 – 15:30 Coffee break

15:30 – 16:00 Breakout sessions reconvene and complete answers to main charge questions

16:00 – 16:15 Breakout leads summarize findings and recommendations in a set of slides and a shared document

16:15 – 17:00 Leads report to the entire workshop (15 min each group including questions)

Day 3 – Thursday, February 27, 2020

07:00 – 08:00 Breakfast

08:00 - 08:15 **Valerio Pascucci, University of Utah & Steering Committee**
Summary of Days 1-2 / Goals of Day 3

08:15 – 08:35 Plenary talk 7, ***Dan Stanzione***, *The University of Texas at Austin*
Providing AI/DL/ML Support in a Large Shared Facility: Perspective from TACC

08:35 – 08:55 Plenary talk 8, ***Eliu Huerta***, *National Center for Supercomputing Applications*
Convergence of AI and HPC for Big-Data Experiments

08:55 – 09:15 Plenary talk 9, **Morris Riedel**, *Juelich Supercomputing Centre*
AI in international Initiatives

09:15 – 10:30 Final breakout sessions: Feedback from community, assign responsibilities, and start drafting the report

10:00 – 10:15 Coffee break

10:15 – 10:45 Continue final breakout sessions: Commit to writing assignments after workshop

10:45 – 11:15 Keynote talk 3, ***Henry Kautz*****, NSF,** Director, Division of Information and Intelligent Systems (CISE/IIS), National Science Foundation
An NSF Vision for AI Computational Infrastructure: From Tensor Clouds to the Edge

11:15 – 11:30 Presentation of draft summary findings and recommendations (steering committee)

11:30 – 12:00 Open discussion and feedback

12:00 – 12:15 Closing remarks, assignments for writing the report

Appendix C

Steering Committee Bios

Dr. Ilkay Altintas is the Chief Data Science Officer at the San Diego Supercomputer Center (SDSC), UC San Diego, where she is also the Founder and Director for the Workflows for Data Science Center of Excellence as well as the WIFIRE Lab, and a Fellow of the Halicioglu Data Science Institute (HDSI). In her various roles and projects, she leads collaborative multi-disciplinary teams with a research objective to deliver impactful results through making computational data science work more reusable, programmable, scalable, responsible and reproducible. Her Ph.D. degree is from the University of Amsterdam with an emphasis on provenance of workflow-driven collaborative science. Among the awards she has received are the 2015 IEEE TCSC Award for Excellence in Scalable Computing for Early Career Researchers and the 2017 ACM SIGHPC Emerging Woman Leader in Technical Computing Award.

Dr. Ian Foster is the Arthur Holly Compton Distinguished Service Professor of Computer Science at the University of Chicago and a Distinguished Fellow and Senior Scientist at Argonne National Laboratory, two wonderful institutions where science and scholarship are cherished, and the importance of new forms of instrumentation as enablers of discovery is appreciated. Dr. Foster is also affiliated with the Department of Computer Science, Mathematics and Computer Science Division, and Institute for Molecular Engineering at the University and Argonne. Dr. Foster interests include building software systems and deploying services that solve problems in the sciences.

Dr. José A.B. Fortes is the AT&T Eminent Scholar and Professor of Electrical and Computer Engineering at the University of Florida where he founded and is the Director of the Advanced Computing and Information Systems Laboratory. Among other current projects, he is leading the development of the cyberinfrastructure of the NSF-funded iDigBio project, is the principal investigator and Steering Committee Chair of the NSF-funded CENTRA initiative and is a principal or co-principal investigator of several other NSF projects (including HuMaIN and PRAGMA). José Fortes is a Fellow of the IEEE and a Fellow of the AAAS.

Dr. Xiaohui (Helen) Gu is a full professor in the Department of Computer Science at the North Carolina State University. She received her PhD degree in 2004 and MS degree in 2001 from the Department of Computer Science, University of Illinois at Urbana-Champaign. She received her BS degree in computer science from Peking University, Beijing, China in 1999. She was a research staff member at IBM T. J. Watson Research Center, Hawthorne, New York, between 2004 and 2007. Dr. Gu received ILLIAC fellowship, David J. Kuck Best Master Thesis Award, and Saburo Muroga Fellowship from University of Illinois at Urbana-Champaign. She also received the IBM Invention Achievement Awards in 2004, 2006, and 2007. She has filed 9 patents, and has published more than 80 research papers in international journals and major peer-reviewed conference proceedings.

Dr. Salim Hariri is a professor and University of Arizona site director of the NSF-funded Center for Cloud and Autonomic Computing. He founded the IEEE/ACM International Symposium on High Performance Distributed Computing, or HPDC, and is the co-founder of the IEEE/ACM International Conference on Cloud and Autonomic Computing. Professor Hariri serves as editor-in-chief of the scientific journal Cluster Computing, which presents "research and applications in parallel processing, distributed computing systems and computer networks." Additionally, he co-authored three books on autonomic computing, parallel and distributed computing, and edited Active Middleware Services, a collection of papers from the second annual AMS workshop published by Kluwer in 2000.

Dr. Valerio Pascucci is the Inaugural John R. Parks Endowed Chair of the University of Utah and the founding Director of the Center for Extreme Data Management Analysis and Visualization (CEDMAV) of the University of Utah. Valerio is also a Faculty of the Scientific Computing and Imaging Institute, a Professor of the School of Computing, University of Utah, and a Laboratory Fellow, of PNNL. Before joining the University of Utah, Valerio was the Data Analysis Group Leader of the Center for Applied Scientific Computing at Lawrence Livermore National Laboratory, and an Adjunct Professor of Computer Science at the University of California Davis. Valerio's research interests include Big Data management and analytics, progressive multi-resolution techniques in scientific visualization, discrete topology, geometric compression, computer graphics, computational geometry, geometric programming, and solid modeling. Valerio is the coauthor of more than two hundred refereed journal and conference papers and is an Associate Editor of the IEEE Transactions on Visualization and Computer Graphics.

Dr. Dan Stanzione is Associate Vice President for Research at The University of Texas at Austin since 2018 and Executive Director of the Texas Advanced Computing Center (TACC) since 2014, is a nationally recognized leader in high performance computing. He is the principal investigator

(PI) for several projects including a multimillion-dollar National Science Foundation (NSF) grant to acquire and deploy Frontera, which is the fastest supercomputer at a US university. Stanzione is also the PI of TACC's Stampede2 and Wrangler systems, supercomputers for high performance computing and for data-focused applications, respectively. He served for six years as the co-director of CyVerse, a large-scale NSF life sciences cyberinfrastructure in which TACC is a major partner. In addition, Stanzione was a co-principal investigator for TACC's Ranger and Lonestar supercomputers, large-scale NSF systems previously deployed at UT Austin.

Dr. Michela Taufer holds the Jack Dongarra Professorship in High Performance Computing in the Department of Electrical Engineering and Computer Science at the University of Tennessee, Knoxville (UTK). Michela Taufer has a long history of interdisciplinary work with scientists across domains and has been serving as the leading principal investigator on several NSF collaborative projects. Michela Taufer is an ACM Distinguished Scientist, an IEEE Senior Member, and an IBM Fellow.

Dr. Xinyu Zhao is an assistant professor of Mechanical Engineering at University of Connecticut. She is the director of the Computational Thermal Fluids Laboratory and her group is actively working on untangling the complex physical processes in turbulent reacting flows using high-performance computing. Xinyu Zhao is the recipient of the AFOSR Young Investigator Award in 2017 and the Irvin Glassman Young Investigator Award from the Eastern States Section of the Combustion Institute in 2020.

Appendix D

Developing a Roadmap for the Next Generation of Smart Cyberinfrastructure

February 25-27, 2020, Hyatt Regency Crystal City, VA,

Call for White Papers

This NSF workshop (http://smartci2020.org/) aims to develop an understanding of the state of the art and the research needs in using scientific cyberinfrastructure powered by artificial intelligence (AI), machine learning (ML), and other smart technologies, as well as using these smart technologies to manage cyberinfrastructure efficiently, with an emphasis on NSF-supported cyberinfrastructure. The scale and priority of both research activities and opportunities that build on smart technologies call for the urgent and coherent development of smart cyberinfrastructures that enables accelerated progress, shared outcomes, high-quality workforce training, and operational excellence. Moreover, the shared research vision should allow for the identification of the most crucial gaps in existing capabilities that are least likely to be filled by industry or NSF or other agencies.

The expected outcomes of this workshop include a cohesive overview of the state of the art in AI, ML, and other smart technologies in the context of several cyberinfrastructure focus areas, including:

- Accelerate AI/ML algorithms on NSF-supported cyberinfrastructure
- Efficiently use NSF-supported cyberinfrastructure through AI/ML technologies
- Provide easy and productive access to AI/ML tools by a broader and diverse scientific community of domestic and international stakeholders
- Explore and assess cyberinfrastructure requirements for AI/ML-based scientific applications

To achieve the above-stated goals and outcomes, this workshop will include plenary presentations, panel discussions, breakout sessions, and lightning talks from contributed whitepapers. Accordingly, we invite you to submit short whitepapers (one- or two-page papers) that address one or more of the topics listed above. The whitepapers should provide your long-term vision and recommendations addressing key aspects that enable the development of a smart cyberinfrastructure, including, among others, the following issues:

- What are the main open research challenges and/or promising research directions in AI/ML and other smart technologies that can positively impact future NSF-supported cyberinfrastructure?

- What computational testbeds for community research can accelerate innovation and lead to future deployments?
- What innovations are needed in the cyberinfrastructure of the future?
- How can the community promote and implement workforce development and training to facilitate the support of cyberinfrastructure?
- What tools and techniques are needed to support cyberinfrastructure-enabled science and engineering?

The topics and questions listed here above are not meant to be comprehensive nor do we expect each white paper to touch all of them. Within the scope of the workshop, you should feel free to develop specific topics and subtopics as well as formulate your own questions. Good ideas about potential lines of important research, emerging new types of relevant technology, or discussions of notable technology gaps are welcome. Putting such ideas in the context of your own experience or the experience of your research community is especially valuable. We will select as many white papers as possible for short presentations (four to five slides; presentations no more than 10 min each) by the authors as lightning talks. The whitepapers can be published as part of the workshop report. The final workshop report will be used to share findings and recommendations with NSF, the scientific communities, and the public.

For full consideration for workshop presentations, send your contributions to Valerio Pascucci (pascucci.valerio@gmail.com) by February 15. We are looking forward to seeing you at the workshop.

Appendix E

White Papers

Artificial Intelligence Accelerated Cyberinfrastructrure as a Research Multiplier at Small and Midsized Institutions

Mike Austin[1,2], Jim Lawson[1,2], Andrew Evans[1], Andrea Elledge[1], and Adrian Del Maestro[1,3]

[1]Vermont Advanced Computer Core, University of Vermont, Burlington, VT, 05405

[2]Enterprise Technology Services, University of Vermont, Burlington, VT, 05405

[3]Department of Physics, University of Vermont, Burlington, VT, 05405

The increasing complexity of scientific data-driven workflows has led to an evolution in the types of tools both employed and demanded by interdisciplinary researchers. A survey of queue wait times and user needs at the University of Vermont Advanced Computing Core (VACC) identified an unmet demand for GPU supercomputing. This gap was recently addressed through the NSF Major Research Infrastructure (MRI) program via the design and deployment of a special purpose device delivering 80 NVIDIA V100 GPUs with a custom-built NVMe over fabric filesystem.

Adoption of the device by advanced users in our cestommunity was rapid, and high utilization developed within months. However, we identified a substantial portion of users that were interested in employing machine learning in their research, but have minimal previous experience with high performance computing and were unable to transition their scientific workflows to GPUs. Thus, the efficient and broad utilization of NSF-supported cyberinfrastructure requires coordinated and simultaneous investment in hardware, research computing facilitation, and community tools that can be deployed for training purposes. We have identified a number of areas where such efforts can have a multiplying effect on research, training, and workforce development.

1. Training workshops that provide a low-level introduction to scientific programming and data analysis that include curricular material on machine learning frameworks with domain-agnostic examples. These should be broadly advertised, especially to those outside of traditional STEM disciplines, and highlight that previous programming experience is not necessary for attendance.

2. Deployment of web-based interfaces (such as Open OnDemand [1]) to research computing assets with environments pre-configured for artificial intelligence applications including Jupyter notebooks. We have seen broad utilization of these technologies by researchers, and high demand for deployment in the classroom by

faculty teaching courses with a computational component, requiring little to no software installation on student laptops.

3. GPU virtualization technologies (e.g. NVIDIA vCompute Server) that enable GPU sharing can lead to higher utilization of expensive physical resources, especially for widely used off-the-shelf applications in materials science and chemistry with speedups reliant on a combination of CPU and GPU, that may not be able to exercise all of the processing capability of a single V100.

4. Campus and region-wide coordination of networking, compute, and storage infrastructure to ensure successful implementation of next-generation imaging applications such as cryogenic electron microscopy and light sheet fluorescence microscopy. There is currently a disconnect between the managers and users of these facilities and the cyberinfrastructure expertise required for sustainability.

The procurement and implementation of specialized cyberinfrastructure with advanced artificial intelligence capabilities is playing an increasing and fundamental role in the university research ecosystem. Enhanced planning and support, especially around training and research software development, is required to ensure broad utilization of smart technologies.

[1] D. Hudak, D. Johnson, A. Chalker, J. Nicklas, E. Franz, T. Dockendorf, and B. L. McMichael. "Open OnDemand: A web-based client portal for HPC centers." J. Open Src. Soft. 3, **622** (2018). https://dx.doi.org/10.21105/joss.00622

AI4IO: A Suite of AI-based tools for capturing IO patterns

Michael Wyatt[1], Stephen Herbein[2], Kathleen Shoga[2], Todd Gamblin[2], Michela Taufer[1]

[1]University of Tennessee Knoxville [2]Lawrence Livermore National Laboratory

1. Problem overview and community needs

Resource managers like SLURM manage nodes and on-node resources, like processors and GPUs but fail to consider the wider set of resources in HPC systems, including the Parallel File System (PFS). The PFS can be a major bottleneck for HPC applications on large-scale systems: peak IO bandwidth is limited by power consumption constraints; concurrent executions compete for IO bandwidth (both on a machine and between different machines connected to the same PFS); and the rate of growth in PFSes to store the data is outpaced by the growth of HPC systems to generate data. Ultimately, HPC application jobs exhibit loss in performance and, in many cases, fail due to contention associated with partial or total loss in the PFSes' IO bandwidth.

Because there is an intrinsic complexity in predicting the occurrence of IO contention patterns (and failures) to access PFSes, current resource managers run short in accurate predictions. Consequently, resource managers often do not make efficient use of HPC resources when dealing with an heterogeneous ensemble of HPC jobs with a broad range of IO usage. When dealing with IO-intensive applications, in the best scenarios, the applications waste scarce node-hours by running their jobs slowly on congested resources; in the worst scenario, entire application allocations are wasted when their jobs fail because of the IO do not complete. Jobs that do very little IO may not be affected at all. If we can infer information about jobs, such as expected runtime and IO usage and which jobs are most affected by IO contention, we can inject the knowledge into the resource managers and ultimately mitigate the loss in performance for the impacted applications.

The community is in need of tools that provide HPC stakeholders with a priori knowledge for managing the wider set of HPC resources. These tools must be able to: (1) identify and exploit temporal and system-specific patterns in resource management; (2) augment resource managers to be resource-aware; and (3) avoid inconveniencing HPC users. Machine Learning (ML) based tools are a promising solution because: (1) ML can capture the trends unique to individual systems or utilities and leverage this for making predictions; (2) ML predictions can be pipelined into HPC resource managers to make schedulers resource-aware; and (3) ML can use the data that is already submitted to HPC machines as input and require no additional information from users.

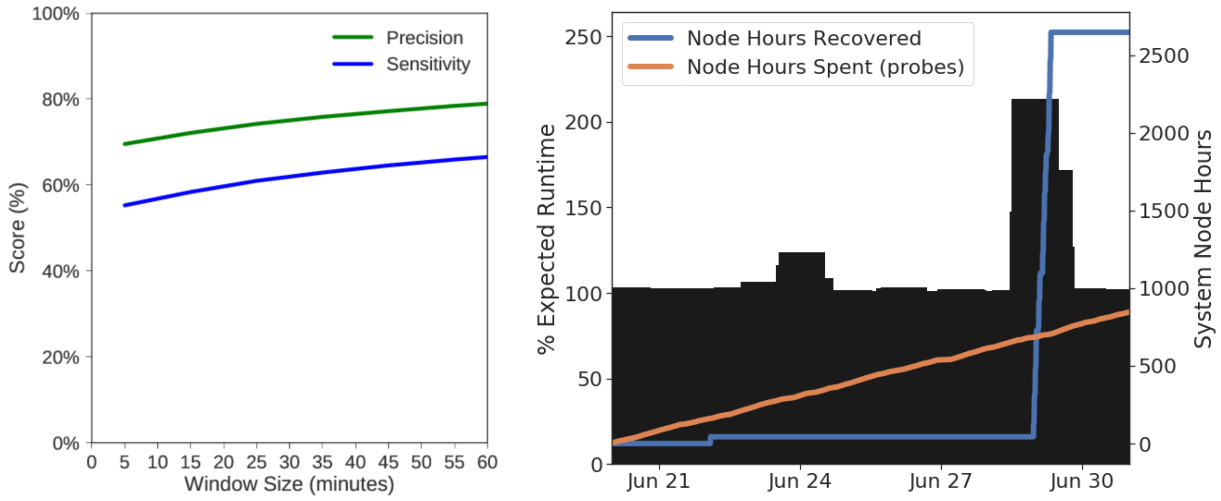
2. Our approach to tackle the problem

We are tackling the overall problem by designing AI4IO, a suite of ML-based tools that capture trends in historical IO data and make predictions on IO patterns that resource managers can use to improve the overall system resource utilization. Our suite currently consists of two tools: PRIONN and CanarIO.

These two tools approach the problem orthogonally to provide complimentary predictions: PRIONN works to **prevent IO contention** by providing resource managers with accurate predictions of individual job runtime and IO usage. We leverage PRIONN predictions to avoid scheduling IO-intensive jobs together, a common cause of IO resource contention and degradation. CanarIO works to **detect and mitigate IO contentions** by predicting the effects of IO degradation in real-time. We leverage CanarIO predictions to detect when IO degradation is occurring and which jobs are IO-sensitive and IO-resilient. We mitigate the contention effect by replacing IO-sensitive jobs with IO-resilient jobs in the execution queue during IO degradation, ultimately recovering wasted system resources.

We evaluate each tool using real HPC data and simulating the execution jobs with a resource manager. Figure 1a shows the precision and sensitivity for which we are able to predict future IO bursts based on PRIONN predictions. We are capable of predicting (and preventing) over 55% of IO bursts within 5 minutes of the actual event. Likewise, Figure 1b shows the node hours spent for data collection to train

CanarIO and the node-hours recovered during an IO degradation event on the system. In just a 10-day window, we observe over 1500 node-hours of saved system resources using predictions from CanarIO.



(a) (b)

Figure 1: Precision and sensitivity for which we are able to predict future IO bursts based on PRIONN predictions (a) and node hours spent for data collection to train CanarIO and the node-hours recovered during an IO degradation event (b).

3. Challenges in using ML for HPC

The accuracy of ML-based tools such as PRIONN and CanarIO heavily relies on the existence of pipelines collecting relevant data and making predictions actionable. In the development of our suite, the major challenge comes from the need for more reliable and accurate data collection infrastructure. Specifically, we observe that the resources being devoted to data collection on HPC machines is insufficient for the deployment of our tools on production machines. PRIONN and CanarIO rely on datasets collected from several sources, including user job scripts, SLURM logs, PFS logs, and system monitoring tools. When one or more of these datasets is unreliable (i.e., missing data or contains incorrect data) the ability to provide accurate predictions to the resource manager diminishes. For this reason, it is necessary to improve data collection infrastructure on HPC machines to support the tools' backbone ML algorithms. In addition to the data-collection infrastructure necessary to support our suite of tools, we observe that innovation of HPC resource managers is necessary. Current resource managers, such as SLURM, do not provide direct and native support for predictions from our tools. Next-generation schedulers should adopt a design model that allows for easy integration and augmentation of the job schedule with tools such as PRIONN and CanarIO.

4. Future directions

We envision future additions to our suite to include tools for monitoring and modeling additional on-node resources, such as GPUs, with constraints for system resources, such as power. Similar to our current tools, these tools will leverage the power of ML to provide predictions to resource managers that enable better scheduling for the wider set of HPC resources under many system- and node-level constraints. In order to develop these tools, new data sources will need to be identified and collected on HPC systems.

An Integrated Approach to Neural Network Design, Training, and Inference

Amir Gholami[1], Michael W. Mahoney[2], Kurt Keutzer[3]
University of California, Berkeley
{amirgh,mahoneymw,keutzer}@berkeley.edu

***Abstract*— Finding the right Neural Network model and training it for a new task requires considerable expertise and extensive computational resources. Moreover, the process often includes ad-hoc rules that do not generalize to different application domains. These issues have limited the applicability and usefulness of DNN models, especially for new learning tasks. This problem is becoming more acute, as datasets and models grow larger, which increases training time, making random/brute force search approaches quickly untenable. In large part, this situation is due to the first-order stochastic gradient descent (SGD) methods that are widely-used for training DNNs. Despite SGD's well-known benefits, vanilla SGD tends to perform poorly, and thus one introduces many (essentially ad-hoc) knobs and hyper-parameters to make it work. It has been found that these hyper-parameters are significantly more sensitive to tuning in large scale training with SGD, and this has impeded effective use of supercomputing systems. Here, we argue that a multi-faceted approach is needed to address these challenges by considering the full stack of neural network architecture design, large scale training, and efficient inference on edge platforms. This requires designing mechanisms to better understand NN training and bridge the gap between theoretical results for optimization, second order methods, and high performance computing.**

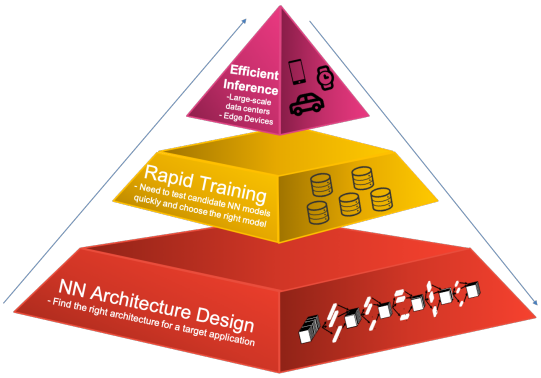


Fig. 1. Reaching the next milestone in large scale machine learning, requires a novel approach by integrating the full stack of designing Neural Network architecture, training, and inference on a target hardware platform. This requires development of new tools/hammers to gain more insight into the problem such as through second-order methods, scalable training methods that are robust to hyper-parameters, and direct integration of inference constraints such as latency/power on a target hardware platform.

I. Background and Significance

Deep Neural Networks (DNNs) have proven to be very effective in diverse applications ranging from semantic segmentation [1, 2] and detection [3, 4] in computer vision to scientific applications such as astronomy [5], climate science [6], and medical image analysis [7, 8]. In these and many other applications of machine learning (ML) and artificial intelligence (AI), finding the right DNN architecture for a particular application and then training a high-quality model requires extensive hyper-parameter tuning and architecture search, often on very large data sets. The delay associated with training DNNs is often the main bottleneck in the design process, and this bottleneck limits the usefulness of DNNs in many applications.

The most straightforward method for accelerating training is to perform the so-called data parallel approach with large batches [9]. However, to efficiently utilize distributed processors, the batch size must grow with the number of processes. In the ideal case, the hope is to decrease the computational time proportional to the increase in batch size, without any drop in generalization quality. However, this is typically not the case; and, in fact, training with large batches often results in poor generalization [10, 11]. It has been found that large batch size training is more likely to converge to the so-called "sharp" local minima, which often do not generalize well. As opposed to this, small batch training has been found to converge to "flatter" local minima that do not have this problem. However, the latter cannot be efficiently scaled to parallel processes.

In order to address these drawbacks, many solutions have been proposed [12–17]. However, these methods either work only for particular models on particular datasets, or they require massive hyper-parameter tuning. While extensive hyper-parameter turning may result in good tables for publications, it is antithetical to the original motivation of using large batch sizes to reduce training time in real applications. This is still an open problem, and it has limited the effective use of supercomputers for these computationally-intensive AI/ML tasks. While one could naively use a supercomputer and perform extensive hyper-parameter sweeps, this is not possible for many large-scale tasks, it is not an efficient use of computational resources, and it often undermines the goal of scientific insight. For example, in our prior work [18, 19] we showed that using large batch size training with SGD leads to diminishing returns. Sample results are shown in Fig. 2, where we show the achieved speed up by increasing batch size for reaching a testing loss threshold. It can be clearly seen that using larger batches does not lead to speed up for a fixed testing loss target. In fact, it has been observed that large batch size training with SGD requires more iterations [20] to recover accuracy which limits the overall speedups. That is even though larger batches can be parallelized more efficiently from a systems perspective, but

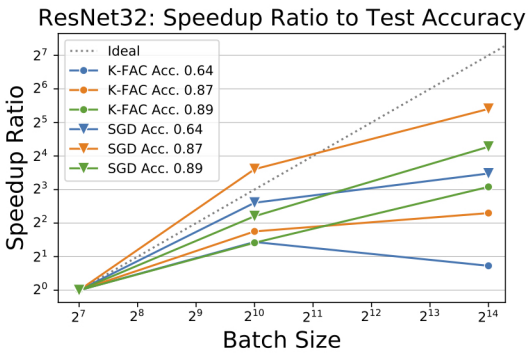


Fig. 2. Speed up to a target testing accuracy versus batch size is shown for both SGD and K-FAC for ResNet32 trained on CIFAR-10. The diminishing returns effect can be seen for both K-FAC (circles) and SGD (triangles). We can clearly see that using larger batches leads to significantly lower speedups as compared to ideal line. For more details please see [19].

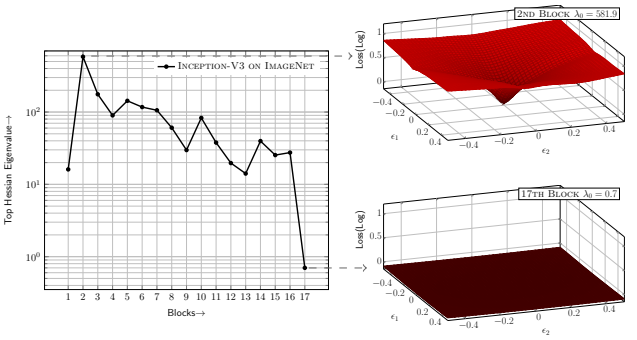


Fig. 3. Top eigenvalue of each individual block of pre-trained Inception-V3 on ImageNet. Note that the magnitudes of eigenvalues of different blocks varies by orders of magnitude [30].

more work/iterations is needed to reach a target accuracy.

The main source for many of these problems stems from the use of the first-order Stochastic Gradient Descent (SGD) algorithm in training DNNs. While SGD has several well-known benefits [21], it is also known to be very sensitive to hyper-parameters such as step size (a.k.a. learning rate), initialization, and momentum. These parameters vary widely from one DNN model to another, sometimes by orders of magnitude. This extreme sensitivity to tuning is exacerbated when performing training with large batches [19]. Therefore, one has to execute many (thousands of) tests to determine the right hyper-parameter values. For example, one of the problems with SGD training is that it uses the same step size for all of the parameters, irrespective of curvature (Hessian) information. Ideally, we want to use larger step sizes for parameters that have small curvature, and vice versa. This is illustrated in Fig. 3, where we show the top Hessian eigenvalue for different layers of Inception-V3 trained on ImageNet dataset. One can clearly see that there is an order of magnitude difference in the associated curvature information. For example, the loss landscape of the last layer of Inception-V3 has very small curvature, which means that a larger step size should be used for those parameters, as opposed to the second layer, which has three orders of magnitude larger curvature, and which thus needs a much smaller step size. Adaptive variants of SGD have been proposed to address this (e.g., AdaGrad and ADAM), but they work only somewhat reliably, and they only work for particular problems.

To address these problems, a multi-faceted approach is needed that can encapsulate the full stack of designing, training, and executing the DNN model on a target hardware platform. All of these stages are interconnected, and focusing only on one area will lead to sub-optimal solutions. This requires designing mechanisms to better understand DNN training and bridging the gap between (i) theoretical results for optimization, (ii) second order methods, and (iii) high performance computing.

Along the first direction, we need to develop a more practical theory for training NNs to enable large scale training of NNs with more robust and interpretable methods. For example, our recent work has developed new theoretical results in the use of higher order optimization methods and in particular their strength and weaknesses as compared to SGD [22–26]. Along the second direction, we have developed the PyHessian framework, which is an open source library that enables fast computation of Hessian spectrum. This includes the top eigenvalue, trace, and even the full Eigenvalue Spectral Density of the Hessian for DNNs [11, 27]. This has lead to new insights in large scale training of DNNs [11, 28], adversarial robustness [29], and development of a novel Hessian AWare Quantization framework [30–33]. The latter has become the state-of-the-art for compressing DNN models through quantization.

Along the third direction, we have developed new methods to scale training by using the so-called integrated parallelism which is based on communication-avoiding algorithms [9]. The algorithm enables distributing the computations by finding optimal partitioning of the data and model, and avoids the problems of large batch size training. Mesh TensorFlow library is a recent work from Google that has used this approach and implemented it in TensorFlow [34].

II. Conclusions

One promising solution to address the challenges associated with large scale training of DNN models is to incorporate recent advances in theory, second-order optimization, and high performance computing. Our recent work, has focused on developing the infrastructure required to address these challenges. In particular, we have developed PyHessian [27, 35] a novel library for second-order based analysis of DNN models, HAWQ [30–33, 36] a library for compressing DNN models for efficient inference at the edge, and integrated parallelism [37] a new algorithm which enables scaling training without changing hyper-parameters.

The next milestone in enabling efficient large scale training of DNN models can be achieved by encapsulating the full stack of training, designing, and executing the DNN model on a target hardware platform. These three phases are intricately related and only focusing on one aspect, will not lead to optimal solutions.

REFERENCES

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[2] B. Wu, A. Wan, X. Yue, and K. Keutzer, "Squeezeseg: Convolutional neural nets with recurrent crf for real-time road-object segmentation from 3d lidar point cloud," in *In Review*, 2017.

[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[4] B. Wu, F. Iandola, P. H. Jin, and K. Keutzer, "Squeezedet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 129–137.

[5] B. Naul, J. S. Bloom, F. Pérez, and S. van der Walt, "A recurrent neural network for classification of unevenly sampled variable stars," *Nature Astronomy*, vol. 2, no. 2, p. 151, 2018.

[6] T. Kurth, S. Treichler, J. Romero, M. Mudigonda, N. Luehr, E. Phillips, A. Mahesh, M. Matheson, J. Deslippe, M. Fatica *et al.*, "Exascale deep learning for climate analytics," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis*. IEEE Press, 2018, p. 51.

[7] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle, "Brain tumor segmentation with deep neural networks," *Medical image analysis*, vol. 35, pp. 18–31, 2017.

[8] A. Mang, S. T. A. Gholami, N. Himthani, S. Subramanian, J. Levitt, M. Azmat, K. Scheufele, M. Mehl, C. Davatzikos, B. Barth, and G. Biros, "SIBIA-GlS: Scalable biophysics-based image analysis for glioma segmentation," *The multimodal brain tumor image segmentation benchmark (BRATS), MICCAI*, 2017.

[9] A. Gholami, A. Azad, P. Jin, K. Keutzer, and A. Buluc, "Integrated model, batch and domain parallelism in training neural networks," *ACM Symposium on Parallelism in Algorithms and Architectures(SPAA'18)*, 2018, [PDF].

[10] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, "On large-batch training for deep learning: Generalization gap and sharp minima," *arXiv preprint arXiv:1609.04836*, 2016.

[11] Z. Yao, A. Gholami, Q. Lei, K. Keutzer, and M. W. Mahoney, "Hessian-based analysis of large batch training and robustness to adversaries," *Advances in Neural Information Processing Systems*, 2018.

[12] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, "Accurate, large minibatch sgd: training imagenet in 1 hour," *arXiv preprint arXiv:1706.02677*, 2017.

[13] Y. You, I. Gitman, and B. Ginsburg, "Scaling sgd batch size to 32k for imagenet training," *arXiv preprint arXiv:1708.03888*, 2017.

[14] A. Devarakonda, M. Naumov, and M. Garland, "Adabatch: Adaptive batch sizes for training deep neural networks," *arXiv preprint arXiv:1712.02029*, 2017.

[15] S. L. Smith, P.-J. Kindermans, and Q. V. Le, "Don't decay the learning rate, increase the batch size," *arXiv preprint arXiv:1711.00489*, 2017.

[16] X. Jia, S. Song, W. He, Y. Wang, H. Rong, F. Zhou, L. Xie, Z. Guo, Y. Yang, L. Yu *et al.*, "Highly scalable deep learning training system with mixed-precision: Training imagenet in four minutes," *arXiv preprint arXiv:1807.11205*, 2018.

[17] Y. You, J. Hseu, C. Ying, J. Demmel, K. Keutzer, and C.-J. Hsieh, "Large-batch training for lstm and beyond," *arXiv preprint arXiv:1901.08256*, 2019.

[18] N. Golmant, N. Vemuri, Z. Yao, V. Feinberg, A. Gholami, K. Rothauge, M. W. Mahoney, and J. Gonzalez, "On the computational inefficiency of large batch sizes for stochastic gradient descent," *CoRR*, vol. abs/1811.12941, 2018. [Online]. Available: http://arxiv.org/abs/1811.12941

[19] L. Ma, G. Montague, J. Ye, Z. Yao, A. Gholami, K. Keutzer, and M. W. Mahoney, "Inefficiency of k-fac for large batch size training," *AAAI'20 (arXiv:1903.06237)*, 2020.

[20] E. Hoffer, I. Hubara, and D. Soudry, "Train longer, generalize better: closing the generalization gap in large batch training of neural networks," in *Advances in Neural Information Processing Systems*, 2017, pp. 1731–1741.

[21] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010)*, 2010, pp. 177–187.

[22] P. Xu, F. Roosta-Khorasan, and M. W. Mahoney, "Second-order optimization for non-convex machine learning: An empirical study," *arXiv preprint arXiv:1708.07827*, 2017.

[23] P. Xu, F. Roosta-Khorasani, and M. W. Mahoney, "Newton-type methods for non-convex optimization under inexact hessian information," *arXiv preprint arXiv:1708.07164*, 2017.

[24] F. Roosta-Khorasani and M. W. Mahoney, "Sub-sampled newton methods i: globally convergent algorithms," *arXiv preprint arXiv:1601.04737*, 2016.

[25] ——, "Sub-sampled newton methods ii: Local convergence rates," *arXiv preprint arXiv:1601.04738*, 2016.

[26] F. Roosta, Y. Liu, P. Xu, and M. W. Mahoney, "Newton-mr: Newton's method without smoothness or convexity," *arXiv preprint arXiv:1810.00303*, 2018.

[27] (2019, Sep.) https://github.com/amirgholami/pyhessian.git.

[28] Z. Yao, A. Gholami, K. Keutzer, and M. Mahoney, "Large batch size training of neural networks with adversarial training and second-order information," *arXiv preprint arXiv:1810.01021*, 2018.

[29] Z. Yao, A. Gholami, P. Xu, K. Keutzer, and M. Mahoney, "Trust region based adversarial attack on neural networks," *Computer Vision and Pattern Recognition (CVPR'19)*, 2018.

[30] Z. Dong, Z. Yao, A. Gholami, M. Mahoney, and K. Keutzer, "Hawq: Hessian aware quantization of neural networks with mixed-precision," *Accepted in International Conference on Computer Vision (ICCV) preprint arXiv:1905.03696*, 2019.

[31] S. Shen, Z. Dong, J. Ye, L. Ma, Z. Yao, A. Gholami, M. W. Mahoney, and K. Keutzer, "Q-bert: Hessian based ultra low precision quantization of bert," *Accepted in AAAI-20 (arXiv:1909.05840)*, 2019.

[32] Z. Dong, Z. Yao, D. Arfeen, Y. Cai, A. Gholami, M. Mahoney, and K. Keutzer, "Trace weighted hessian-aware quantization," *NeurIPS'19 workshop on Beyond First-Order Optimization Methods in Machine Learning*, 2019.

[33] Y. Cai, Z. Yao, Z. Dong, A. Gholami, M. W. Mahoney, and K. Keutzer, "Zeroq: A novel zero shot quantization framework," *arXiv preprint arXiv:2001.00281*, 2020.

[34] N. Shazeer, Y. Cheng, N. Parmar, D. Tran, A. Vaswani, P. Koanantakool, P. Hawkins, H. Lee, M. Hong, C. Young *et al.*, "Mesh-tensorflow: Deep learning for supercomputers," in *Advances in Neural Information Processing Systems*, 2018, pp. 10 414–10 423.

[35] Z. Yao, A. Gholami, K. Keutzer, and M. Mahoney, "PyHessian: Neural Networks through the lens of the Hessian," *under review*, 2019.

[36] (2020, Jan.) https://github.com/amirgholami/zeroq.git.

[37] A. Gholami, K. Kwon, B. Wu, Z. Tai, X. Yue, P. Jin, S. Zhao, and K. Keutzer, "Squeezenext: Hardware-aware neural network design," *Workshop paper in CVPR*, 2018.

White Paper for the NSF Workshop on Smart Cyberinfrastructure, Washington, USA, February 25-27, 2020

Prof. Dr. – Ing. Morris Riedel
Head of Research Group, Juelich Supercomputing Centre, Forschungszentrum Juelich, Juelich, Germany
Professor of AI, School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland
Web: www.morrisriedel.de – Twitter: @MorrisRiedel

Selected European Perspectives and Trends as Top 5 Recommendations for Smart Cyberinfrastructures for AI in the Future

Executive Summary
There is a wide variety of activities of developing 'Cyberinfrastructures in Europe' (i.e., rather known as Research infrastructures) in several application domains such as those driven forward by the European Strategy Forum for Research Infrastructures (ESFRI) [1] that are primarily based on domain-specific use of data, computing, and tools. A key European Artificial Intelligence (AI) – driven effort that can not directly be considered as a 'Cyberinfrastructure', but is rather a 'AI on-demand platform' (AI4EU) [2] shares several ambitious goals with Cyberinfrastructures. This 'top-down' initiated platform is currently in a very early stage of development but aims to inform the AI community about AI news, new AI tools, emerging AI services, and enable the sharing of AI techniques and algorithms between users of the platform and AI-related EU projects. In addition to European activities also national AI activities emerge like the Helmholtz AI initiative [11] funded by the Helmholtz Association in Germany. At the same time, we observe that 'European Cyberinfrastructures' primarily focussed on computing and storage like the Partnership for Advanced Computing in Europe (PRACE) [3] or the European Grid Initiative (EGI) [4] work on satisfying an increasing number of requests related to AI applications. For example, this includes annual PRACE training courses such as 'Parallel and Scalable Machine Learning' [5] or 'Introduction to Deep Learning Model' [6]. The feedback of users in these training courses and the below outlined research activities contributed to lessons learned summarized as top five recommendations below for NSF w.r.t. development of smart cyberinfrastructures for AI in the future.

Despite the momentum in deep learning, users of Cyberinfrastructures in Europe such as PRACE also increasingly use traditional feature engineering approaches like parallel component trees [15], or machine learning techniques like parallel and scalable Support Vector Machines (SVM) [12], or traditional data mining methods like parallel and scalable Density-based spatial clustering of applications with noise (DBSCAN) [13]. Especially for application areas with less labeled samples, this makes sense while we observe more and more availability of data (i.e., often quoted 'big data') but that does not necessarily mean that 'big data' is of high quality or perfectly labeled to be used with machine learning techniques. For cases where large quantities of data are available, Cyberinfrastructures support distributed deep learning training at scale, for example by using Horovod for scaling across many GPUs for AI applications like in remote sensing [14]. Internally, Horovod is using concepts of the Message Passing Interface (MPI). Also, users demand increased number of open datasets for AI applications.

Given the fact that many users of Cyberinfrastructures continue to use it with simulation sciences based on numerical methods based on known physical laws, the increase in using AI workloads with new frameworks becomes a challenge for Cyberinfrastructures leading to more required heterogeneity in the Cyberinfrastructures. Examples include also the use of Apache Spark for AI workloads like using autoencoders [17] or the joint use of AI in physics-informed ML applications [21]. Resource providers need to support a broader set of workloads that investigations in the European DEEP series of projects [16] reveal leading to a modular supercomputing architecture (MSA) [22] implemented in the Juelich Supercomputing Centre (JSC). The modularity will be key for Cyberinfrastructures to support, not only to support traditional simulation sciences and more emerging AI workloads but also to integrate innovative computing aspects (e.g., quantum computing, neuromorphic computing, etc.). We have already started to use an SVM based on the quantum annealing approach [19] to solve an inherent optimization problem in remote sensing applications [18]. Hence, Quantum computing is in reach and need to be part of future 'modular' concepts of Cyberinfrastructures. Another example of the need for modularity are MSA-enabled reinforcement learning environments that, for example, may help to deal with the large complexity of finding the right hyper-parameters in deep learning networks.

Recommendation 1: Enable Application Enabling Process after Cyberinfrastructure Training
After performing Cyberinfrastructure & AI training (i.e., lessons learned from PRACE training), users often raise demand for so-called 'application-enabling' by high level support teams (HLSTs) that can be summarized as supporting AI users in several weeks of adopting new AI techniques within their application domains using Cyberinfrastructures; training and offering 'some core-hours' alone is not enough to ensure an uptake of the Cyberinfrastructure capabilities and offered services & tools.

Recommendation 2: Provide Easy Access and Sharing of Scripts & Data with JupyterLab
PRACE training courses for AI as described above (but also for other topics like Python, MPI, etc.) are increasingly carried out with Jupyter notebooks and JupyterLab and this environment is available also for researchers at the Juelich Supercomputing Centre (JSC) and was presented at ISC2019 last year [10]. Cyberinfrastructures need to support these technologies more broadly to enable easier collaboration in research through the sharing of scripts, Jupyter kernels, or datasets for AI applications. This also lowers the technical boundary for AI end-users that are not familiar with concepts like SSH and could be considered for integration into existing Science Gateway activities in the US. Challenges for Cyberinfrastructure providers need to be tackled like supporting interactive access instead of the traditional batch operations and implement Jupyter kernels on various coherent versions of deep learning tools (e.g., specific versions of Tensorflow/Keras) and underlying hardware libraries (e.g., CUDDN).

Recommendation 3: Contribute with Cyberinfrastructures to International AI Activities
There are excellent examples of international collaboration activities such as the Joint Laboratory for Extreme-Scale Computing (JLESC) [7] or Big Data & Extreme-Scale Computing (BDEC) activities [8] that both are not primarily focussed on AI and rather focus towards developments in Exascale. Hence, in contrast to the AI4EU platform, the Joint Artificial Intelligence and Machine Learning Lab (JoAIML) [9] was created as a spin-off from JLESC as Bottom-Up activity to encourage the exchange of AI-based scripts and datasets via GitLab [9]. Although JoAIML is currently driven forward by the Helmholtz AI at JSC, the University of Iceland (UoIceland), and the National Center for Supercomputing Applications (NCSA), JoAIML is considered to be open to other international members (e.g., other partners from US, Japan, UK, etc.). More recently, discussions emerge how JoAIML can become broader and more visible (e.g., own Web page) for Cyberinfrastructure users in the future.

Recommendation 4: Offer Parallel & Scalable Tools & Techniques at Scale
Despite the fact that Deep Learning tools (e.g., TensorFlow, Keras, PyTorch, etc.) and distributed training frameworks (e.g., using Horovod) are required, it makes sense also to offer also traditional machine learning algorithms as parallel and scalable codes in the Cyberinfrastructure. Not all applications really need cutting-edge deep learning models like RESNET-50, for example, when labeled dataset samples are limited that is observed quite often in science and engineering. For example, SVMs, as desribed above, are robust methods that may lead to better results in some application areas. Reasons are grounded in statistical learning theory and that the number of parameters for SVMs is not much (in contrast to deep learning models) thus enabling better generalization and less risk in overfitting.

Recommendation 5: Modularity of the Cyberinfrastructure with distinct Services & Resources
The modularity of the Cyberinfrastructure will be key for AI applications in the future and is here inspired by the MSA as described above. While computational-intensive model training can be done by using cluster modules (CM) with high single-thread performance CPUs or specific GPUs, less computational-intensive AI model testing and inference can be performed on scalable booster modules (BM). Apache Spark stacks with AI algorithms can be supported by Data Analytics Modules (DAMs) contributing to the fact that for specific AI application problems it is more efficient to use specific devices. Like accelerators are used now for deep learning problems, it can be possible that a quantum device module can be used for the complex solving of optimization problems in AI algorithms. Equally interesting, future resources and services may offer neuromorphic devices or more use of containers (i.e., Docker, Singularity, etc.) for AI applications. Other distinct services of the Cyberinfrastructure should support end-users in a systematic fashion by enabling (instance-aware) Neural Architecture Search (NAS), for example, based on reinforcement learning environments as described in [20]. Finding the right hyper-parameter and parameter-tuning is very computational-intensive, overlaps with AutoML techniques, and is a key problem to be solved for AI communities by Cyberinfrastructures of the future.

References

[1] European Strategy Forum on Research Infrastructures (ESFRI), Online:
https://www.esfri.eu/
[2] AI on Demand Platform for Europe (AI4EU), Online:
https://www.ai4eu.eu/
[3] Partnership for Advanced Computing in Europe (PRACE), Online:
http://www.prace-ri.eu/
[4] European Grid Initiative (EGI), Online:
https://www.egi.eu/
[5] PRACE Tutorial: Parallel and Scalable Learning – Introduction, Online:
http://www.morrisriedel.de/prace-tutorial-parallel-and-scalable-machine-learning-introduction
[6] PRACE Tutorial: Introduction to Deep Learning Models, Online:
http://www.morrisriedel.de/introduction-to-deep-learning-models
[7] Joint Laboratory for Extreme-Scale Computing (JLESC), Online:
https://jlesc.github.io/
[8] Big Data & Extreme-Scale Computing (BDEC), Online:
https://www.exascale.org/bdec/
[9] Joint Artificial Intelligence and Machine Learning Lab (JoAIML), Online:
https://gitlab.version.fz-juelich.de/JoAIML_Lab/workspace/material
[10] Goebbert, J.H., Kreuzer, T., Grosch, A., Lintermann, A., **Riedel, M.**: **Enabling Interactive Supercomputing at JSC Lessons Learned**, in conference proceedings of the International Conference on High-Performance Computing, Springer, Lecture Notes in Computer Science (LNCS) Vol. 11203, June 24-28, 2019, Frankfurt, Germany, Online:
https://www.researchgate.net/publication/330621591_Enabling_Interactive_Supercomputing_at_JSC_Lessons_Learned_ISC_High_Performance_2018_International_Workshops_FrankfurtMain_Germany_June_28_2018_Revised_Selected_Papers
[11] Helmholtz AI Initiative of the Helmholtz Association, Online:
https://www.helmholtz.ai/
[12] Cavallaro, G., **Riedel, M.**, Richerzhagen, M., Benediktsson, J., Plaza, A.: **On Understanding Big Data Impacts in Remotely Sensed Image Classification Using Support Vector Machine Methods**, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS), Vol 9 (10), pp. 1-13, 2015, Online:
https://www.researchgate.net/publication/282524415_On_Understanding_Big_Data_Impacts_in_Remotely_Sensed_Image_Classification_Using_Support_Vector_Machine_Methods
[13] Goetz, M., Bodenstein, C., **Riedel, M.**: **HPDBSCAN – Highly Parallel DBSCAN**, in conference proceedings of ACM/IEEE International Conference for High-Performance Computing, Networking, Storage, and Analysis (SC 2015), Machine Learning in HPC Environments (MLHPC 2015) Workshop, November 15-20, 2015, Austin, Texas, USA, Online:
https://www.researchgate.net/publication/301463871_HPDBSCAN_highly_parallel_DBSCAN
[14] Sedona, R., Cavallaro, G., Jitsev, J., Strube, A., **Riedel, M.**, Benediktsson, J.A.: **Remote Sensing Big Data Classification with High Performance Distributed Deep Learning**, Journal of Remote Sensing, Multidisciplinary Digital Publishing Institute (MDPI), Special Issue on Analysis of Big Data in Remote Sensing, 2019, Online:
https://www.researchgate.net/publication/338077024_Remote_Sensing_Big_Data_Classification_with_High_Performance_Distributed_Deep_Learning
[15] Goetz, M., Cavallaro, G., Geraud, T., Book, M., **Riedel, M.**: **Parallel Computation of Component Trees on Distributed Memory Machines**, IEEE Transactions on Parallel and Distributed Systems (TPDS), Vol 29 (11), 2018, Online:
https://www.researchgate.net/publication/325212343_Parallel_Computation_of_Component_Trees_on_Distributed_Memory_Machines
[16] DEEP Series of EU Projects, Online:
https://www.deep-projects.eu/
[17] Haut, J.M., Gallardo, J.A., Paoletti, M.E., Cavallaro, G., Plaza, J., Plaza, A., **Riedel, M.**: **Cloud Deep Networks for Hyperspectral Image Analysis**, IEEE Transactions on Geoscience and Remote Sensing, PP(99):1-17, 2019, Online:
https://www.researchgate.net/publication/335181248_Cloud_Deep_Networks_for_Hyperspectral_Image_Analysis
[18] G. Cavallaro, D. Willsch, M. Willsch, K. Michielsen, and **M. Riedel**: **Approching Remote Sensing Image Classification with Ensembles of Support Vector Machines on the D-Wave Quantum Annealer**, in the IEEE International Geoscience and Remote Sensing Symposium, 2020 (submitted)
[19] D. Willsch, M. Willsch, H. De Raedt and K. Michielsen, **'Support Vector Machines on the D-Wave Quantum Annealer'** 2019, Online:
http://dx.doi.org/10.1016/j.cpc.2019.107006
[20] **Morris Riedel**, ,'**Neural Architecture Search with Reinforcement Learning**' Invited Talk, 5th International Summer School on Big Data and Machine Learning, Technical University of Dresden, Dresden, Germany, Online:
http://www.morrisriedel.de/neural-architecture-search-with-reinforcement-learning
[21] Mathis Bode, Michael Gauding, Zehu Lian, Dominik Denker, Marco Davidovic, Konstantin Kleinheinz, Jenia Jitsev, Heinz Pitsch, '**Using Physics-Informed Super-Resolution Generative Adversarial Networks for Subgrid Modeling in Turbulent Reactive Flows**', Online:
https://www.researchgate.net/publication/337560077_Using_Physics-Informed_Super-Resolution_Generative_Adversarial_Networks_for_Subgrid_Modeling_in_Turbulent_Reactive_Flows
[22] E. Suarez, N. Eicker, Th. Lippert, ,'**Modular Supercomputing Architecture: From Idea to Production**', Book, Contemporary High Performance Computing, Online:
https://www.researchgate.net/publication/334264090_Modular_Supercomputing_Architecture_From_Idea_to_Production

Accelerated AI for Edge Computing

Maryam Rahnemoonfar

Associate Professor of AI
Computer Vision and Remote Sensing Laboratory (Bina lab), UMBC
maryam@umbc.edu

1 Introduction

The volume and complexity of multidimensional signals have been growing exponentially. Analysis of the big datasets collected by various sensors remains a significant challenge for scientists and analysts. There is also a growing need for real-time analysis of data on Edge in many applications including emergency response, health care, surveillance, and cybersecurity. The faster one can harness insights from data, the greater the benefit in reducing costs, saving human lives, and increasing efficiency will be. While traditional analyses provide some insights into the data, the complexity, scale, and multi-disciplinary nature of the data necessitate advanced intelligent solutions. The success of recent data analytics technique based on deep learning have facilitated progress in a variety of tasks but they command significant computational complexity and require the availability of a large amount of labeled data for training.

2 Challenges and needs

Despite all recent advances of deep neural networks, there are currently several challenges; firstly, success of recent deep learning approaches for a variety of vision-based tasks highly depends on the availability of a large amount of labeled data for training. In many real-world applications, such as natural disasters and health-care, manually labeling images requires a significant amount of domain experts' time that could otherwise be spent on high-level scientific discovery. Secondly, most successful network architectures (such as VGG-net[1], Resnet[2]) have improved the performance of various vision tasks at the expense of significantly increased computational complexity and as a result, they need several days for training on graphics processing units

(GPUs). In many applications, fast analysis of data is vital. For example, after natural disasters, if one can reduce the initial response by one day, one can reduce the entire recovery by a thousand days [3]. Another example is in health-care; if one detects an elderly fall immediately, one can save his/her life. Thirdly, several studies have shown that even a limited amount of noise and perturbation greatly affects the performance of most deep learning techniques [4, 5]. Finally, the interpretability of black box representations of deep learning has long been the Achilles' heel of deep learning. It achieves superior performance at various tasks at the cost of low interpretability.

Critical need: There is a critical need in research community and industry to develop deep learning algorithms that are fast, unsupervised, less computational, and less sensitive to the noise.

3 Our vision

To address the aforementioned challenges in deep learning in general, we propose to devise new methodologies in time-frequency domain based on multi-resolution analysis. This domain-transformed deep learning approach has the potential to solve the issues of low training speed, and susceptibility to noise, as well as the need for large amounts of labeled data. Wavelets provide several advantages when performing deep learning operations in that domain, including sparse representation, multi-resolution, and space-frequency locality.

References

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[3] Robin Murphy, "Ted talks: These robots come to the rescue after a disaster," 2015, https://www.youtube.com/watch?v=wG4RnDNWtJo, Last accessed on 2018-7-14.

[4] S. Dodge and L. Karam, "Understanding how image quality affects deep neural networks," in *Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on.* IEEE, 2016, pp. 1–6.

[5] A. Fawzi, O. Fawzi, and P. Frossard, "Analysis of classifiers' robustness to adversarial perturbations," *Machine Learning*, vol. 107, no. 3, pp. 481–508, 2018.

[6] M. Rahnemoonfar, "Fast solution of phase unwrapping partial differential equation using wavelets," *Electronic Journal of Differential Equations*, vol. 23, pp. 119–129, 2016.

[7] G. Beylkin, "On the representation of operators in bases of compactly supported wavelets," *SIAM Journal on Numerical Analysis*, vol. 29, no. 6, pp. 1716–1740, 1992.

[8] G. Beylkin, R. Coifman, and V. Rokhlin, "Fast wavelet transforms and numerical algorithms i," *Communications on pure and applied mathematics*, vol. 44, no. 2, pp. 141–183, 1991.

[9] I. Drori and D. Lischinski, "Fast multiresolution image operations in the wavelet domain," *IEEE transactions on visualization and computer graphics*, vol. 9, no. 3, pp. 395–411, 2003.

[10] Y. Liu, M. M. Cheng, X. Hu, K. Wang, and X. Bai, "Richer convolutional features for edge detection," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 2017, pp. 5872–5881.

[11] J. Pont-Tuset, P. Arbelaez, J. T. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping for image segmentation and object proposal generation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 1, pp. 128–140, 2017.

[12] T. C. Wang, M. Y. Liu, J. Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," *arXiv preprint arXiv:1711.11585*, 2017.

AI-assisted HPC simulations and application to astrophysics and cosmology

Tiziana Di Matteo, Rupert Croft and Yueying Ni (Carnegie Mellon), Simeon Bird (UC Riverside), Yin Li (Flatiron), Yu Feng (Google)

Dealing with finite resolution and resources. As with all HPC simulations of physical systems, cosmological simulations of galaxy formation [1, 8] are limited by insufficient computational resources. It is possible to directly follow the equations of gravity and gas dynamics either at high resolution in a small volume or at low resolution in a large volume. However, key processes of interest often result from the interplay of large scale environments and small scale physics. Drawing from the ongoing revolution in AI (specifically, Deep Learning, DL, e.g., [3]) it is becoming possible to bridge this gap from large to small scales. Neural Networks (NNs) have been developed that can learn from high resolution image data [2, 5], and then make accurate super resolution versions of different low resolution images. We are beginning to apply these techniques [6] to physical modeling in HPC simulations. The ideas are generally applicable, and our own use case is cosmological hydrodynamics, leading to super resolution modeling of galaxy and black hole formation.

Training Generative Networks. The heart of these AI-assisted simulations are *Generative Adversarial Networks (GANs)* [4, 7]. A GAN is a class of DL system in which two NNs contest against each other in a game: the generative network generates candidates while the discriminative network evaluates them. Given a training set, this technique learns to generate new data with the same characteristics as the training set. GANs can be used to generate entirely new data from initially random inputs, or in the case of super resolution simulations we use low resolution full hydrodynamics data as input to the generative network and produce data statistically consistent with high resolution simulations below the resolution scale. Training sets are generated by running simulations at different native resolutions, and results from different physics codes can be combined.

The path to hybrid AI/hydrodynamic modeling. Three different approaches are being tried, in increasing order of complexity (i) enhancement of simulated galaxy images; (ii) post processing augmentation of simulated three dimensional gas, dark matter and stellar density fields; and (iii) training and super resolution modeling inside running simulations ("on the fly") which allows for causal feedback between the generated and physically simulated scales. .

On-chip AI: The suites of super resolution models will make use of on-chip AI instructions becoming available in HPC, such as Intel's AVX-512 Vector Neural Network Instruction set available on the Cascade Lake processors used in the NSF's Frontera. Such technology holds great promise to study problems where large scales and high resolution are closely linked. In the case of astrophysics these super resolution simulations will bridge the gap in scales between the formation of stars, planets and black holes and the dark matter and dark energy that dominate the large-scale structure of the Universe. Our understanding of both ends of this range — from small to large — will increase as a result.

Impact on NSF Cyberinfrastructure We will be running AI algorithms inside advanced physics simulations, with training and use happening in real time, as the AI will affect the running simulation and learn from it at the same time. This level of coupling between AI and direct physical modeling is unprecedented and should lead to advances in the ways that AI can be used. Flexible tightly coupled systems of GPUs and CPUs with

on-chip NN would be ideal resources for both NN training farms, and on the fly AI/fluid dynamics runs. If this work fulfills its promise it could revolutionize the use of AI-enabled chips in exascale high performance computing.

REFERENCES

[1] R. Croft, T. Di Matteo, N. Khandai, and Yu Feng. Petascale cosmology: Simulations of structure formation. *Computing in Science & Engineering*, 17(02):40–46, mar 2015.

[2] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *CoRR*, abs/1501.00092, 2015.

[3] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[5] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. *CoRR*, abs/1609.04802, 2016.

[6] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[7] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[8] Mark Vogelsberger, Federico Marinacci, Paul Torrey, and Ewald Puchwein. Cosmological simulations of galaxy formation, 2019.

White Paper Submitted to NSF Workshop on Smart Cyberinfrastructure, Feb 2020

The Role of Machine Learning in Scientific Workflow Management on Distributed Cyberinfrastructure

Anirban Mandal (anirban@renci.org), Ewa Deelman (deelman@isi.edu)
10 February, 2020

Scientific workflows are key to today's computational science, enabling the definition and execution of complex applications in heterogeneous and often distributed cyberinfrastructure (CI). There are several current challenges in managing scientific workflows in distributed systems: composing the workflows, provisioning the resources for the workflows, and executing the workflows efficiently and reliably. Promising machine learning (ML) techniques can be applied to meet these challenges thereby enhancing the current workflow management system capabilities [1]. We foresee that as the ML field progresses, the automation provided by workflow management systems will greatly increase and result in significant improvements in scientific productivity and reproducibility. We herein describe a few representative possible applications of ML technologies that can enable efficient and reliable use of NSF-supported CI by scientific workflows.

Integrity Introspection for Scientific Workflows using ML: Data-driven science workflows often suffer from unintentional data integrity errors when executing on distributed national scale CI. However, today, there is a lack of tools that can collect and analyze integrity-relevant data from workflows and thus, many of these errors go undetected jeopardizing the validity of scientific results. In the context of the IRIS project [2], we are developing methods to automatically detect, diagnose, and pinpoint the source of unintentional integrity anomalies in scientific workflows executing on distributed CI. The approach is to develop an appropriate threat model and incorporate it in an integrity analysis framework that collects workflow and infrastructure data and uses ML algorithms to perform the needed analysis. Our goal is to integrate our solutions into the Pegasus workflow management system [3], which is used by a wide variety of scientific domains. It is also important to engage with science application partners (e.g. gravitational-wave physics, earthquake science, and bioinformatics) to deploy the ML analysis framework for their workflows, and to iteratively fine tune the threat models, ML model training, and ML model validation in a feedback loop.

Testbed Experimentation and ML Model Validation: An important aspect of developing ML based analysis techniques is the appropriate use of testbed infrastructures (NSFCloud [4][5] and other NSF-supported testbeds, e.g. ExoGENI [6]) to simulate realistic infrastructure conditions and error scenarios to train the ML models. For example, we are currently simulating aspects of the Open Science Grid [7] data distribution system in a testbed scenario to research on introspection and diagnosis of data integrity errors. The goal is to build an analysis framework that is powered by novel ML-based methods developed through experimentation in a controlled testbed, and then validated in and made broadly available on NSF production CI. Future NSF-supported infrastructures like the FABRIC [8] mid-scale research infrastructure can also be brought to bear to develop ML models for workflows with realistic production scenarios.

Performance Analysis for Workflows: ML techniques can also be utilized to address many of the challenges faced in managing scientific workflows in distributed systems. Several ML techniques are being used today to analyze the behavior of workflows at various levels of abstraction (workflow, task, and infrastructure) using different processing modalities (online and offline) and techniques to train the ML models [1]. Although there are promising initial results in using ML algorithms for scheduling, anomaly

Contact author: Anirban Mandal (anirban@renci.org)

detection and provisioning resources for efficient and reliable workflow executions, the community is just at the beginning of exploring ML techniques in the scientific workflow space. Significant research advances need to be made in applications of ML technologies to the area of workflows to truly automate the analysis of the workflow behavior, understand the sources of anomalies, and make adaptation decisions to efficiently support the entire workflow lifecycle.

References

[1] E. Deelman, A. Mandal, M. Jiang and R. Sakellariou, "The role of machine learning in scientific workflows," in *International Journal of High Performance Computing Applications*, doi: https://doi.org/10.1177/1094342019852127, May 2019.

[2] IRIS website. https://sites.google.com/view/iris-nsf/home

[3] E. Deelman, K. Vahi, G. Juve, M. Rynge, S. Callaghan, P. J. Maechling, R. Mayani, W. Chen, R. Ferreira da Silva, M. Livny, and K. Wenger, "Pegasus: a workflow management system for science automation." *Future Generation Computer Systems*, vol. 46, pp. 17–35, 2015.

[4] NSF Chameleon Cloud. https://chameleoncloud.org/

[5] NSF CloudLab. https://cloudlab.us/

[6] I. Baldin, J. Chase, Y. Xin, A. Mandal, P. Ruth, C. Castillo, V. Orlikowski, C. Heermann, J. Mills. "ExoGENI: A Multi-Domain Infrastructure-as-a-Service Testbed." *The GENI Book*, pp. 279--315, 2016.

[7] Open Science Grid. https://opensciencegrid.org/

[8] FABRIC. https://whatisfabric.net/

Contact author: Anirban Mandal (anirban@renci.org)

SUSTAINABLE CYBERINFRASTRUCTURES

WHITE PAPER

Mark A. Asch
University of Picardy,
Amiens, FRANCE.
EXDCI2 Project
mark.asch@u-picardie.fr

Thierry Bidot
Neovia-Innovation
Paris, FRANCE
EXDCI2 Project
thierry.bidot@neovia-innovation.eu

François Bodin
University of Rennes
Rennes, FRANCE
EXDCI2 Project
francois.bodin@irisa.fr

Maike Gilliot
ETP4HPC/TERATEC
Bruyères-le-Chatel, FRANCE
EXDCI2 Project
maike.gilliot@teratec.fr

Laurent Morin
INRIA
Rennes, FRANCE
laurent.morin@inria.fr

Recent hardware and software advances have motivated the development of a transcontinuum digital infrastructures concept to account for the convergence of data and compute capabilities *across* the digital continuum, which goes from instruments and sensors at the edge, through wireless and network channels, to HPC and cloud facilities et the center. This concept is not in a straight line from the past efforts and a paradigm change is needed: we will have to design systems encompassing hundreds of billions of cores distributed over scientific instruments, IoT, supercomputers and Cloud systems through Lan, Wlan and 5G networks.

Pushed by massive deployments of compute and storage capabilities at the edge, we require new system design to accommodate the ecosystem change to be expected in the coming decades (environmental and technological) and horizontally integrate the different actors. The new demands and challenges that combine data and compute, distributed across the continuum, and the maintenance and resource efficiencies, are pushing for drastically increased *software and hardware sustainability*. Furthermore, the need to provide high-level cybersecurity is profoundly changing the game. Efficiency and resilience will have to reach levels never achieved so far, while taking into account the intrinsic distributed and heterogeneous nature of the continuum. In addition, the question of dealing with such high volumes of data needs to be faced, and quality versus quantity will become unavoidable. These considerations will spread over all components. Long-lifetime hardware devices will have to be reconfigurable, modular, and self-aware in order to be operational over extended periods. Algorithm efficiency will need to be drastically pushed up (e.g. more efficient AI). Management and deployment of large-scale application workflows will have to be adapted, or invented. Network protocols will have to offer better control over the data logistics, etc.

It is widely recognized that AI will play a central role in these extreme-scale, continuum infrastructures. This will occur at three levels: (1) AI for Digital Infrastructure, (2) Digital Infrastructure for AI, and (3) AI for Science, Industry and Societal Challenges. The first addresses how AI can pilot and monitor the continuum and in so doing provide solutions to the points listed in the previous paragraph. The second treats the question of re-designing the cyberinfrastructure to efficiently deal with data analysis and machine learning, which means tuning of data access, I/O, and low precision arithmetic. The last deals with the ever-increasing needs to exploit AI techniques for extreme-scale, combining Data and Compute through the interpretation and coupling of computing results, measurements and observations (eg. Digital Twins in extreme earth modelling, combining climate models with satellite data and on-ground sensors).

The overall objective is to target high TRL solutions (7 and more), based on horizontal synergies between all the concerned digital infrastructure technologies: HPC, Big Data, Machine Learning, IoT, 5G, cybersecurity, processor technology and robotics. All of these components of the digital infrastructure will together be able to address the huge societal challenges and sustainable development goals by mobilizing their amazing potential all the way across the continuum.

References

[1] Asch, Moore, et al. Big data and extreme-scale computing: Pathways to Convergence-Toward a shaping strategy for a future software and data ecosystem for scientific inquiry. *Int. J. HPC and Applications*, Vol. 3, No. 4, pages 435–479. 2018.

[2] ETP4HPC. *Strategic Research Agenda For High-Performance Computing in Europe, January 2020 - European HPC Research Priorities 2021 – 2024.* Edited and published by ETP4HPC (European Technology Platform for High Performance Computing) 2020. `https://www.etp4hpc.eu/sra.html`

Developing and Accelerating Predictive Models for Predicting Relapse of Pediatric Oncology patients using Smart Cyberinfrastructure

Mauricio Ferrato, Brad Atmiller, Erin Crowgey, Karl Franke, Sunita Chandrasekaran
Affiliation: University of Delaware, Nemours/Alfred I. duPont Hospital for Children

Goal: To apply ML/AI algorithms to clinical and genomic data of nearly 200 Leukemia patients to predict their relapse potential.

Dataset: The TARGET (Therapeutically Applicable Research To Generate Effective Treatments) from National Cancer Institute (NCI) initiative provides comprehensive molecular characterization to the greater research community. TARGET has applied this approach to multiple types of cancer so far. Our focus is on Acute Lymphoblastic Leukemia (ALL), a cancer of white blood cells. One half of our data is each patient's Clinical Annotations; each patient's demographic data (like age and race) and medical data (like White Blood Cell count and Minimal Residual Disease) throughout their treatment. The other half is their genetic profile. One major benefit of using TARGET genetic data is that it is already pre-processed and clustered into gene groups with levels of expression for each. Instead of low-level gene data and nucleotides, we get to work with a spreadsheet that has 20,000 gene groups for each of the 200 patients.

Project Focus: Our project focuses on using both clinical and genetic data of Acute Lymphoblastic Leukemia (ALL) patients to determine molecular changes that drive childhood cancers. We aim to find what mix of feature selection techniques and classification models provide the most accurate and generalizable results.

Approach: The first important step for any data science project is **pre-processing,** getting the data into a form that machine learning algorithms will recognize and can process. The primary data structure we use to hold our data is DataFrames from the Python Data Analysis Library (pandas). From there we clean up the data by removing columns that are missing too many values, fill in the rest of the absent value (using averages or placeholders), split categorical variables into a binary representation, and scale numeric columns to a consistent range.

The second step is **feature selection**. One of the major pieces of our research is feature selection; how to best determine which features will be most useful in predicting relapse. The most accurate learning algorithms are those which are fed high-quality data. One possible formalization is that a good feature set contains features which are highly correlated with the class, yet uncorrelated with each other. Specifically determining which genes out of the 20,000 will be most useful is quite difficult. There are a few different routes. One route is performing statistical analysis to find which variables are most closely correlated with relapse. Chi-squared statistic, correlation, information gain, Symmetrical uncertainty, ReliefF statistic, and a host of other forms of analysis are possible. One of the main analyses we've used is the ratio of expression for each gene between relapse and non-relapse patients. If we take the datascience route, there are several techniques we could employ to perform feature selection. Techniques could be Market Basket Analysis, PCA and autoencoder among others.

The third step is **classification.** Once we have handled missing data and selected features, we to determine the type of classier to use to arrive at predictive and accurate models for the data under study. We will explore artificial neural networks, and ensemble models based on random splits, bootstrapping and cross validation. We will gather accuracy, specificity and sensitivity metrics and analyze how each of the models arrive at these metrics. If the above ensemble algorithms are not useful for the dataset under study, we use another ensemble method called the gradient boosting algorithm for the given dataset. Gradient boosting is one of these ensemble techniques that uses weak learners, like small decision trees called decision stumps, to learn in an iterative process. While using these ensemble approaches, we want to be cautious and observant that the model does not pick a random sample every time for multiple iterations. We also need to explore the applicability of our findings on a larger dataset or dataset that is different to the one under study.

Research Challenges: As we can see, there are more than a handful of techniques one could use to perform feature selection or classification. Such an exploration is also very closely tied to the input dataset and the volume of the same, often termed as "Big Data" . In order to arrive at an acceptable % for accuracy, sensitivity and specificity (ROC curve for example) by the community, we are often bound to employing permutations and combinations of techniques. However, this is a very time-consuming approach. We need **modular workflows**.
Secondly, we clearly need **better software** to be able to use ML/AI solutions. We need interoperability between frameworks such that programmers are not re-inventing the wheel when they have to move between architectures. Given for example Habana relies on a base architecture for both training and inference but optimizes the designs for different workload as opposed to another route where a company manufactures fundamentally different architectures for training and inference. The challenge is how do you target these architectures for the case under study – building models for predictive oncology?
Thirdly, while disruptive hardware is fantastic for exploration, IMHO the community has not come to a consensus on the architectures we need for training and inference. Like mentioned above, sometimes the base architecture is the same and sometimes they are not. A better analysis or study of which is a better route would be very useful. To that end, we need ML/DL benchmark suite like SPEC CPU/HPG. MLPerf is something we have been referring to, however how comprehensive is the suite?

Workforce Development and Training: Jupyter notebooks are one way to go that can be adopted by the next-generation workforce in order to use the smart cyberinfrastructure without needing to get into the nitty-gritties of the workflow itself.

Tools and Techniques: We need "modular" workflows that can walk researchers through techniques and strategies employed for a given dataset. Scripts are OK but they are not helpful when the scientists have to test more than a handful of tools and techniques before arriving at the right combinations of tools for a given dataset.

A Smart Cyberinfrastructure should use AI/ML to Reduce its Carbon Footprint and Negative Environmental Impacts

Andrew A Chien, University of Chicago
White Paper for the 2020 NSF Smart Cyberinfrastructure Workshop

Computing is one of the largest consumers of electrical power and in most parts of the world, the fastest growing consumer. While the largest fraction of this growth is due to the growth of cloud computing [Nature19], the scientific computing community's carbon footprint has grown rapidly, with a recent estimate for the Top 500 at 3 million metric tons of carbon per year. While the growth is due in part to the end of Moore's Law and the rising appreciation of the value of scientific computing, and there is a growing recognition that expanding use of AI/ML is a major contributor to this rapid increase [Amodei18,Hao19].

We believe that the NSF's Smart Cyberinfrastructure should use AI/ML to reduce its carbon footprint and negative environmental impacts. Specifically,

- AI/ML should be used to schedule workloads both at centers and at the edge to minimize the carbon impact of cyberinfrastructure computing resources, choosing location, time of day, and more to consume the least damaging electric power
- AI/ML should be used to manage and plan cyberinfrastructure resources so as to achieve scientific ends with the minimum quantity of e-waste, maximizing the useful lifetime of equipment and scheduling computations on the resources that minimize the consumption of electric power

In addition, the NSF's Smart Cyberinfrastructure should make aggressive use of renewable energy, exploiting breakthrough insights that show that batch and throughput scientific computing can be delivered economically with zero or near zero-carbon approaches [ZCCloud19,Yang16]. This insight has several implications and opportunities:

- NSF should exploit new types of renewable-energy models (stranded power), and new models of operation to achieve zero-carbon operation for large-scale computing resources, doing so can also reduce the power cost of operating these resources
- NSF Smart Cyberinfrastructure should provide national and international thought leadership and operational leadership for both government, educational, and even commercial community to reduce the carbon impact of computing
- NSF Smart Cyberinfrastructure adoption of these renewable based models represents the **opportunity** to maximize the ability to use compute-intensive AI/ML techniques for scientific and social good as well as to optimize scientific computing resource management

References

[Nature18] Jones, N. How to stop data centres from gobbling up the world's electricity. *Nature* (Sept. 12, 2018).

[Amodei18] Amodei and Hernandex, “Ai and compute,” openai.com/blog/ai-and- compute/, May 2018, 10x growth per year.

[Hao19] K. Hao, “Training a single ai model can emit as much carbon as five cars in their lifetimes,” Technology Review, June 2019.

[ZCCloud19] Chien, Zero-Carbon Cloud Project, http://zccloud.cs.uchicago.edu/

[Yang16] Yang, F. and Chien, A.A. ZCCloud: Exploring Wasted Green Power for High-Performance Computing, IPDPS, May 2016.

Artificial Intelligence Accelerated Cyberinfrastructrure as a Research Multiplier at Small and Midsized Institutions

Mike Austin[1,2], Jim Lawson[1,2], Andrew Evans[1], Andrea Elledge[1], and Adrian Del Maestro[1,3]

[1]Vermont Advanced Computer Core, University of Vermont, Burlington, VT, 05405

[2]Enterprise Technology Services, University of Vermont, Burlington, VT, 05405

[3]Department of Physics, University of Vermont, Burlington, VT, 05405

The increasing complexity of scientific data-driven workflows has led to an evolution in the types of tools both employed and demanded by interdisciplinary researchers. A survey of queue wait times and user needs at the University of Vermont Advanced Computing Core (VACC) identified an unmet demand for GPU supercomputing. This gap was recently addressed through the NSF Major Research Infrastructure (MRI) program via the design and deployment of a special purpose device delivering 80 NVIDIA V100 GPUs with a custom-built NVMe over fabric filesystem.

Adoption of the device by advanced users in our cestommunity was rapid, and high utilization developed within months. However, we identified a substantial portion of users that were interested in employing machine learning in their research, but have minimal previous experience with high performance computing and were unable to transition their scientific workflows to GPUs. Thus, the efficient and broad utilization of NSF-supported cyberinfrastructure requires coordinated and simultaneous investment in hardware, research computing facilitation, and community tools that can be deployed for training purposes. We have identified a number of areas where such efforts can have a multiplying effect on research, training, and workforce development.

1. Training workshops that provide a low-level introduction to scientific programming and data analysis that include curricular material on machine learning frameworks with domain-agnostic examples. These should be broadly advertised, especially to those outside of traditional STEM disciplines, and highlight that previous programming experience is not necessary for attendance.

2. Deployment of web-based interfaces (such as Open OnDemand [1]) to research computing assets with environments pre-configured for artificial intelligence applications including Jupyter notebooks. We have seen broad utilization of these technologies by researchers, and high demand for deployment in the classroom by

faculty teaching courses with a computational component, requiring little to no software installation on student laptops.

3. GPU virtualization technologies (e.g. NVIDIA vCompute Server) that enable GPU sharing can lead to higher utilization of expensive physical resources, especially for widely used off-the-shelf applications in materials science and chemistry with speedups reliant on a combination of CPU and GPU, that may not be able to exercise all of the processing capability of a single V100.

4. Campus and region-wide coordination of networking, compute, and storage infrastructure to ensure successful implementation of next-generation imaging applications such as cryogenic electron microscopy and light sheet fluorescence microscopy. There is currently a disconnect between the managers and users of these facilities and the cyberinfrastructure expertise required for sustainability.

The procurement and implementation of specialized cyberinfrastructure with advanced artificial intelligence capabilities is playing an increasing and fundamental role in the university research ecosystem. Enhanced planning and support, especially around training and research software development, is required to ensure broad utilization of smart technologies.

[1] D. Hudak, D. Johnson, A. Chalker, J. Nicklas, E. Franz, T. Dockendorf, and B. L. McMichael. "Open OnDemand: A web-based client portal for HPC centers." J. Open Src. Soft. 3, **622** (2018). https://dx.doi.org/10.21105/joss.00622

Finding the Lost:
True Dissemination of Large Data through Efficient and Standardized System Design

Brian Summa, Assistant Professor, Department of Computer Science, Tulane University

Datasets many gigabytes to terabytes in size are being produced daily by scientists throughout the world. Examples include large simulations of physical phenomena being run on DOE supercomputers or large NSF supported infrastructures; large microscopy scans from NSF or NIH supported projects; or even climate models run at the National Center of Atmospheric Research. While crucial to the scientists who produce them, just a small number of these large datasets could have a transformative impact for others, if the data can be shared simply and easily with researchers. For instance, consider computer vision databases that collect, organize, and share the large number of photographs available online. Databases like ImageNet [1] have fostered huge leaps in vision and machine learning (16k citations in 10 years). Plug-ins even exist to query this database directly in interactive data analytics pipelines (e.g. Jupiter notebooks). Now consider that just a few hundred (~500) datasets from high-resolution, digital microscopy contain approximately the same number of pixels as the entire ImageNet [2].

Imagine the potential if such data could be made widely and easily available to researchers-at-large. This begs the obvious question on what would be necessary for people to construct *ImageNet-like* databases using their large datasets. The unfortunate truth is that even sharing a single large dataset can be a daunting task for researchers. Typically these data repositories are maintained by individual scientists or groups. These ad-hoc repositories are especially challenging when datasets get large. For instance, the repositories are often *links* to full-resolution raw data. In this scenario, datasets cannot be easily transferred, stored, or processed given their sheer size. Moreover, interactive queries are not supported, despite interactive analytics often being the key for novel scientific insights in new or complex phenomena. Alternatively, scalable, interactive queries can be provided through more advanced large data systems [3], although these systems are often highly-specialized, ad-hoc deployments with no standardization for data access across different platforms. In total, meaningful and wide dissemination of large data is not currently well-supported

The only solution to sharing these datasets is through large data systems that are easy to deploy, query, and scale. To this end, the ad-hoc deployments, divergent designs, non-standardized data access, and layering of heuristics that currently plague these systems must end. Work must be completed to model and standardize these many approaches. Although, this is a difficult task. For instance, systems for large data need to leverage (concurrently) several if not all of the following accelerators: high performance backends; preprocessing data while loading or idle; pre-caching or pre-fetching of data; or stream processing during data movement. Moreover, often scalable systems will need to support heterogeneous, distributed computing and storage resources. This leads to a highly complex interplay of many moving parts. Although, as this list above also shows there is justification that such standardization is possible given the many commonalities in design between these ad-hoc systems. Therefore, I argue that such standardization is not only necessary, but entirely possible given the proper support and effort. With such work, the rich datasets that are currently being lost to researchers-at-large due to their size can finally be widely disseminated.

[1] J. Deng et al. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

[2] C. Mercan et al. Multi-instance multi-label learning for multi-class classification of whole slide breast histopathology images. *IEEE Transactions on Medical Imaging*, 37(1):316–325, 2017.

[3] Bethel, E. Wes, Hank Childs, and Charles Hansen, eds. *High performance visualization: Enabling extreme-scale scientific insight. CRC Press, 2012.*

Multi-AI Capable Cyberinfrastructure

Glenn.Ricart@{us-ignite.org, Utah.edu}

White Paper for Workshop on Next Generation of Smart Cyberinfrastructure

February 2020

This White Paper will provide a viewpoint on (1) how scientific cyberinfrastructure will need to evolve to better support a wider range of Artificial Intelligence (AI) techniques and methodologies (which I call Multi-AI-Capable), and (2) example uses for Multi-AI-Capable Cyberinfrastructure in scientific discovery.

Scientific Cyberinfrastructure to Better Support a Wider range of AI Techniques & Methodologies

Existing NSFCloud infrastructures such as CloudLab and Chameleon as well as commercial cloud providers such as Azure and Amazon Web Services (AWS) are rapidly adding hardware optimized for machine learning and classification of patterns. These include fast, parallel, and narrow graphics processing units (GPUs), tensor processing units (TPUs), and smart network interface controllers (smart NICs) among others. There continues to be rapid evolution of hardware optimized for specific uses and keeping up with application-based requirements is a key challenge for both NSF Cloud and commercial providers.

Most the newest and fastest supercomputers are also incorporating architectures designed to facilitate AI workloads. The Japanese AI Bridging Cloud Infrastructure (ABCI) is #8 on the current TOP500 list, but this list ranks by the traditional High Performance Linpack (HPL) benchmark and doesn't measure true AI capabilities. Both next-generation Chinese and European supercomputers will have internal connectivity and augmented and adjunct processors friendly to multi-dimensional AI problems. Two examples are the Vector Neural Network Instructions (VNNI) in current and future Intel Xeon CPUs, and continuous new takes on tensor co-processors.

To date, most AI algorithms and cyberinfrastructure rely on just two related AI techniques: machine learning of patterns and classification based on those patterns. This white paper argues not that these are unimportant—indeed, they are critically important—but that future AI applications will also rely on additional AI-related cyberinfrastruture capabilities as well. Generically, I call this broader set of capabilities "Multi-AI Capable Cyberinfrastructure."

Multi-AI Capable Cyberinfrastructure will better support:

- Traditional symbolic AI reasoning and ontological reasoning
- Being able to explain why learning and classification networks are constructed the way they are
- Game-theoretical extrapolations and predictions (leading to better machine and human decision-making)
- AI-driven automatic big data exploration and discovery and testing these results against independent datasets
- Evolutionary and genetically-driven AI capabilities

- Quantum computing support for better probabilistic AI applications (Quantum wave function superpositioning might be used to represent probabilistic certainties (aka “uncertainties”) in AI algorithms)

Each of these bullets deserves a paragraph to explore their importance and possible approaches to cyberinfrastructure to better support them. But let’s leave that to a workshop report that covers those (and others) on which we have a degree of agreement.

Multi-AI-Capable Cyberinfrastructure in Scientific Discovery

There would appear to be a very long list of possible applications for Multi-AI-Capable Cyberinfrastructure in Scientific Discovery. Before listing some of these, let’s just note that we’re omitting big classes of applications which have already received significant attention such as military applications, autonomous vehicles, digital assistants, traffic management, social media, interpreting medical imaging, financial investing, etc. And we’re not claiming these are already solved; just that we’re likely to make more of an NSF difference by focusing on some important issues in science.

How might Multi-AI Capable Cyberinfrastucture better support science and engineering?

1. Intelligent agents to facilitate use of the Scientific Method. An AI-based digital lab assistant could help support documentation of the environment, hypotheses, helping with experimental design to maintain desired power in the results, accurately recording experiments (including those run on Multi-AI Capable Cyberinfrastructure), and aiding reproducibility and replicability. “Reproducibility and Replicability in Science” was the title of a Congressionally-required, NSF-commissioned report by the National Academies of Science and Engineering released May 7th,. Using Multi-AI Capable Cyberinfrastructure to better design and document scientific hypotheses and experiments could have great value across all NSF directorates.

2. Creation, testing, and tuning of scientific models / simulations / emulations / cyberphysical systems. The Wikipedia article on Artificial Intelligence notes that “Leading AI textbooks define the field as the study of "intelligent agents": any device that perceives its environment and takes actions that maximize its chance of successfully achieving its goals.” (See English Wikipedia for references.) I would suggest that models (and simulations and emulations and self-adjusting cyberphysical systems) are key to many areas of science. A grand challenge might be a National Library of AI-driven Scientific Models. Many AI-driven topic-specific models already exist. But they are few compared to the number of topic-specific models carefully constructed by knowledgeable scientists and engineers. Multi-AI techniques could be the key to greatly expanding the breadth and number of useful models and helping them evolve to match new data. Which brings us to…

3. Discovery in data that goes beyond simple patterns and looks for multi-step explanations and interactions. How are time-varying complex patterns related to each other? Can Multi-AI-Capable Cyberinfrastructure help with model verification? Model-driven prediction? Model tuning? Model explanation?

4. Cybersecurity research is a constantly evolving subject that might be viewed as based on game theory. Multi-AI-Capable Cyberinfrastructure designed for game theory applications could help produce significant new results in cybersecurity.

5. Both student-directed and Socratic learning digital assistants. Today's dominant uses of linear textbooks, syllabus-based lectures, and educational projects work for some but not all learners. The National Science Foundation has an important interest in science and technology education that might better be achieved for some students with experience-inspired student-directed learning. Can Multi-AI-Capable Cyberinfrastructure re-structure learning materials on the fly to meet the needs of a learner's question? Similarly, the Socratic Method is often very effective for some students but not widely used for teaching science and engineering. How could a Multi-AI-Capable Cyberinfrastructure ask probing questions, understand the student's thinking based on their response, and then ask additional questions that lead the student to the answers they seek? (Another grand challenge.)

6. Smart and Connected Communities research could widely benefit from Multi-AI-Capable Cyberinfrastructure. A quick listing of example applications that would benefit from Multi-AI Capable techniques includes:

    a. Anticipating human needs (e.g. public safety, transportation, health)
    b. Consumer-driven AI to make it an even fight with marketing-driven AI (information asymmetry)
    c. Continuous health monitoring
    d. Robotic companions for seniors
    e. Advising citizens on changing existential habits (uses of energy, water, etc.)
    f. Micro transit coordination
    g. Micro grid coordination
    h. Countering social media addiction
    i. Ethical AI; where AI intersects with people, we can't ignore the ethics

White paper on **Collaborations to Enable Transnational Smart Cyberinfrastructure Research, Applications and Workforce** by José Fortes, University of Florida.[i]

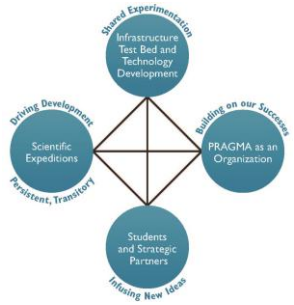
"…envisions vibrant partnerships among academia, government laboratories and industry, including international entities, for the development and stewardship of sustainable CI services that can enhance productivity and accelerate innovation in science and engineering." *In "Transforming Science Through Cyberinfrastructure: NSF's Blueprint for a National Cyberinfrastructure Ecosystem for Science and Engineering in the 21st Century"*

Broadly construed, Artificial Intelligence (AI) goes beyond algorithms and technology to include cultural, ethical and socio-technical issues. For AI to generalize in a global context it is essential to engage with transnational experts and users who are best positioned to understand those issues as well as to interface and integrate with technologies and user policies that differ across countries. In many scientific domains (e.g. biodiversity and environmental monitoring) data may be physically bound to specific locations and expressed in local languages. Therefore, learning may have to occur in a distributed fashion and research collaborations may require access and use of resources across different national cyberinfrastructures. These cyberinfrastructures typically use different hardware/software technologies, different management policies and external network connectivity. In this context, several cyberinfrastructure challenges are faced, including:

- No framework for establishing, maintaining and executing transnational collaborative projects.
- Inexistence of trusted networks of researchers upon which to build such a framework and a community with shared research goals that need transnational cyberinfrastructure.
- Incomplete easy-to-use technical solutions for connecting cyberinfrastructures and overlaying trustworthy transnational systems across them.
- No shared understanding of local requirements for solutions to global problems.
- Very small number of researchers who are aware of variations across countries of socio-technical issues affecting shared/connected cyberinfrastructure and the problems that require its usage.
- Insufficient training opportunities for workforce capable of contributing to transnational cyberinfrastructure and/or its applications.

Successful solutions to the above-listed challenges exist for big science in specific domains such as physics and astronomy, as illustrated by recent well-publicized successes in observation of gravitational waves, detection of the Higgs particle and imaging of a black-hole. However, when dealing with the long-tail of (medium and small scale) science, most current solutions are ad-hoc, clunky, transient and built on personal links. There is a need for systematic, effective, persistent and structured approaches to the creation, execution and sustenance of collaborations for the development, usage and application of shareable transnational cyberinfrastructure. The NSF-funded PRAGMA and CENTRA projects (http://www.pragma-grid.net/ and http://www.globalcentra.org/) exemplify ongoing efforts to address this need and may provide a basis for international collaborations on smart cyberinfrastructure.

PRAGMA's activities



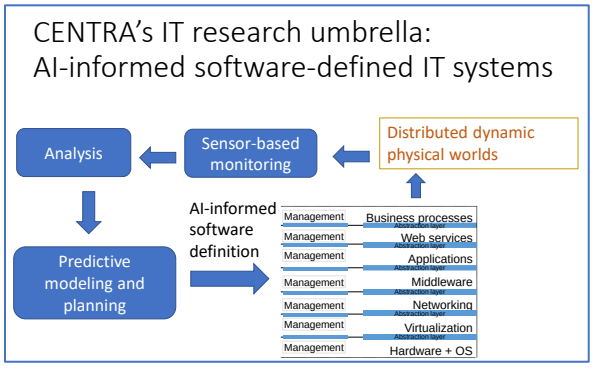
PRAGMA has been building an international, distributed community of practice, primarily around the Pacific Rim, for technology and approaches that support the long tail of science, namely enabling small- to medium-sized international groups, to make rapid progress in conducting research and education by providing and developing international, experimental cyberinfrastructure. To realize this mission, PRAGMA's current activities include four interrelated activities:

- Fostering international "scientific expeditions" by forging teams of domain scientists and cyberinfrastructure researchers who develop and test information technologies that are needed to solve specific scientific questions and create usable, international-scale, cyber environments; There are three current expeditions:
    - Biodiversity: Understanding adaption in extreme environments.
    - Limnology: Predicting lake eutrophication and training the next generation of lake scientists who are part of the Global Lakes Ecological Observatory Network (GLEON), an international organization, that

grew out of an early PRAGMA experiment, whose mission is to understand, predict and communicate the role and response of lakes in a changing global environment.
    - ENT: Developing an experimental network testbed for experimenting with software-defined networks and monitoring impacts of choices.
- Developing and improving a grassroots, international cyberinfrastructure for testing, computer-science insight and, advancing scientific applications by sharing resources, expertise and software;
- Infusing new ideas by developing young researchers who gain experience in cross-border science and by extending engagements with strategic partners;
- Building and enhancing the essential people-to-people trust and organization developed through regular, face-to-face meetings - a core foundation of PRAGMA's success.

CENTRA currently integrates participants from universities and research centers in the US, Korea, Japan, Portugal and Taiwan who meet annually and host visiting researchers involved in collaborations. CENTRA's mission is:

- To enable research on cyberinfrastructure technologies needed to address transnational societal needs in domains that include but are not limited to disaster management, environmental modeling and smart and connected communities; and to advance the science needed to design and use these technologies to build effective and efficient transnational IT systems. Collaborative projects emerge organically from interactions that take place within the CENTRA research network and annual meetings.
- Train the next generation of junior researchers and innovators to work in transnational settings through
    - Convening experts in workshops and webinars to communicate or identify solutions to key problems.
    - Providing junior researchers an immersive experience in collaborative, multidisciplinary teams that address the transnational problems (including stays at international sites involved in collaborations).
    - Facilitating the use and development of prototypes and testbeds to demonstrate solutions.



Communities like PRAGMA and CENTRA advance targeted science areas through developing, applying, and lowering the barriers to use of critical information technologies. Funding is highly leveraged, with each country funding its own researchers and cyberinfrastructure. These communities form the backbone people-network to create a unique, coordinated initiative to provide opportunities for cyberinfrastructure workforce training. Therefore, we have the opportunity to create multi-site programs that (a) identify promising researchers in early to mid-career, (b) gives them multiple rotational global experiences in labs run by leading global researchers, and (c) homes them strategically in their respective national research institutions. Over time, this would create a corps of international researchers who will advance global research effort through multidisciplinary and multinational collaboration. The guiding principle is to create strategically structured problem-based learning activities that enhance student skill sets with long-term engagement for students as they move from undergraduate to graduate stages within PRAGMA/CENTRA and among multiple institutions. This coordinated, multi-lateral approach, which is expected to provide additional value to both students and institutions, is a key difference when comparing this program to other bilateral approaches and provides different types of benefits for students, institutions, and companies.

---

[i] Acknowledgement and Disclaimer: This paper is based on experience and discussions with many people, in the US and internationally, in PRAGMA and CENTRA over several years. The US parts of PRAGMA and CENTRA are funded by grants OAC 1234983 and ACI 1550126 from the National Science Foundation. In particular, I want to recognize Peter Arzberger, Jason Haga, Prapaporn Rattanatamrong, Glenn Ricart and Shava Smallen for discussions about international research-collaboration experiences for students. Errors and any lack of clarity are mine alone.

Social Science Research Towards Widespread and Sustainable Smart Cyberinfrastructures

Kerk F. Kee
College of Media & Communication
Texas Tech University
Lubbock, TX, USA
kerk.kee@ttu.edu

***Abstract*—This paper suggests that in order to build smart cyberinfrastructures towards being widespread and sustainable, the efforts would benefit from complementary insights from social science research. It proposes five areas of social science research – user adoption and systemic diffusion, diverse workforce development, co-production between developers and users, thriving online communities, as well as inclusive and ethical infrastructures. The paper concludes with two recommendations for the NSF to consider – creating a 'communication management plan' requirement under the broader impacts criterion, and the (re)development of a social science research funding program under the Office of Advanced Cyberinfrastructure.**

Keywords—social science research, sustainable cyberinfrastructure, technology adoption, widespread innovation

I. Introduction

The 2020 vision of making cyberinfrastructure (CI) 'smart' with artificial intelligence (AI) and machine learning (ML) is exciting. Smart CI will take e-science, computational social science, and digital humanities to new heights. Over the last two decades or so, CI has been steadily developing and maturing. This can be seen in the strategic investment of the TeraGrid in 2001, the publication of the Atkins Report in 2003, the initial establishment of the NSF Office of Cyberinfrastructure (OCI) in 2005-2006, the iteration of the TeraGrid into XSEDE in 2011, the reinventions of OCI into the Division of Advanced Cyberinfrastructure in 2012-2013 and the Office of Advanced Cyberinfrastructure in 2015-2016.

CI development started as a technical phenomenon. However, CI has also been argued as a socio-technical system [1]. This recognition is evident in the establishment of the Virtual Organizations as Sociotechnical Systems (VOSS) program under NSF OCI, which funded social science research on CI and technologies roughly between 2006 to 2013. With AI and ML being key components of smart CI, understanding the human and social dimensions is more important than ever.

This paper proposes five social science research areas that can help generate insights towards making smart CI widespread and sustainable over the long-term. The premise here is that the NSF can maximize its investments in smart CI when CI also achieves widespread adoption and has a thriving ecosystem around it. This outcome requires research insights and practical strategies derived from the social sciences, and social science research can complement CI's technical foundation.

This material is partially based upon work supported by the US National Science Foundation (NSF) under awards #1322305 and #1453864.

II. Social Science Towards Building Widespread and Sustainable Smart Cyberinfrastructures

A. User Adoption and Systemic Diffusion

A smart CI without users is void of data, activities, and outcomes. It is like a mall without customers, impressive in structures but without impacts on the economy. Moving from traditional experimental, theoretical, local, and small-scale research towards large-scale computational research with big data and empowered by smart CI requires a fundamental shift in the way researchers do their work. Meaningful use of smart CI begins with adoption by individual users.

User adoption occurs at the individual level, and for smart CI to have the greatest impact on research and its applications for scientific breakthroughs, the US economy, citizens' well-being, national security, etc., user adoption needs to expand into systemic diffusion. 'Diffusion' refers to the widespread of smart CI as a platform innovation in the overall research community [2]. Such a diffusion of innovations that transformed the research community have been seen in the adoption of personal computers in the 1980s and the Internet in the 1990s. Smart CI is initiating a new wave of transformations.

The need to conduct social science research to understand user adoption and systemic diffusion is the first strategic move towards becoming widespread. Le et al. [3] advanced the argument to debunk the romantic ideal of – "*If you build it, they will come*," which they termed a 'fallacy'. Without strategic promotions, smart CI may struggle with attracting users and attracting them quickly. Social science research on technology adoption and innovation diffusion [2] can support such an effort.

B. Diverse Workforce Development

To continue building on the metaphor of the mall for smart CI – a buzzing mall also needs shopkeepers, cleaning crew, and security guards to ensure the smooth operation of businesses for customers. For smart CI to be successfully implemented, understanding how to develop and maintain a diverse workforce is critical. Social science research has shown that teams with diverse members are generally 'smarter' – better at generating new ideas and increasing innovations in projects [4].

Smart CI needs a diverse and 'smart' workforce to facilitate user training and user support, following widespread adoption and diffusion. What are the effective strategies in training and

support? This remains an open research question. CI users do not usually come into a training workshop like a traditional student in the classroom – a blank slate to be given generic information about a topic. Most CI users come with specific questions and needs. Social science research on problem-based learning can optimize training. CI trainers also require having effective communication skills, beyond technical knowledge.

C. Co-Productions between Developers and Users

Generic user training is sufficient when users only need CI technologies 'straight out of the box'. However, for many projects, users need custom-made technologies for specific research questions, methodologies, and datasets. In these cases, developers will work with users to identify their needs and design custom-made technologies. The co-production process is complex because many users do not fully know what is possible and/or what they really need. It is through multiple meetings between developers and users that they identify the solutions together. Effective co-productions will require research insights from user-centered design and human-computer interactions.

To better identify users' needs, and the requirements to prototype and test new tools through multiple iterations, another research topic is identifying a software development methodology, such as agile software development (ASD), for effective co-productions. The argument here is not to promote ASD, as ASD and its variations have practical limitations. The point here is to highlight the need to study the social dimensions in co-production in the development of smart CI.

D. Thriving Online Communities

Smart CI can develop thriving online communities of researchers across time and space. With all the pieces interwoven into a robust CI, researchers within and across domains can share and integrate data now and over a long period of time, to do large-scale and longitudinal research otherwise not possible. However, understanding virtual organizations, interdisciplinary collaborations, international partnerships, and online community building are all complex human endeavors that would benefit from the social sciences, such as organizational communication, organizational sociology, and computer supported cooperative work. The goal is to create and sustain a thriving ecosystem to carry smart CI forward.

E. Inclusive and Ethical Infrastructures

Finally, a smart CI is also an ethical CI. It pays attention to who are included and excluded in the process of advancing CI research. Leigh Star and Geoff Bowker in science and technology studies wrote about the concept of the 'installed base' [5]. They caution against the activities in laying the foundation of any infrastructure, such that an 'installed base' with specific values and priorities are built into an infrastructure, creating a situation where future technologies and communities not fitting this base will be excluded. In the early phases of infrastructural developments, it is likely that decisions are made based on what makes sense for the immediate problems. It is likely that no groups are intentionally trying to exclude others in the future of the infrastructure being built at the moment. However, the argument here is to raise the awareness of how the decisions made together today will have future implications, and CI developers would be wise to keep the 'long now' [6] in mind.

Furthermore, for smart CI to be optimal, it requires continuous gathering and processing of data, including data with sensitive information. This is especially true in the case of biomedical research, for example. Also, university committees, such as the Institutional Review Boards (IRBs), may have strict policies on governing the collection and longitudinal use of sensitive data for research. In the case of medical research with smart CI, AI algorithms and ML techniques may rise to becoming 'deterministic'. How can patients' wishes and their human values not be override by AI and ML in the name of being 'smart' is an important ethical question. It would be wise for smart CI developers to be mindful of the human ethical dimensions of CI, and/or work with social scientists and critical scholars who can help them attend to this particular challenge.

III. Conclusion

This paper proposes five specific areas in social science research that can help build smart, widespread, and sustainable CI towards transformation of research across domains, locations, and time. Recounting the argument advanced by Le et al. [2] – "*If you build it, users may not come*". In order to address this concern, the NSF can consider two suggestions. First, similar to the requirement introduced in 2012 for proposals to include a data management plan, the NSF could consider adding a similar requirement (or an optional supplemental) for a 'communication management plan' under the 'Broader Impacts' criterion. Such a plan would encourage PIs to be more thoughtful about promoting user adoption and systemic diffusion, cultivating a diverse workforce, facilitating co-production between developers and users, building thriving online communities, and/or designing inclusive and ethical infrastructures in their smart CI projects. These social science topics are inherently 'communication' in nature. Second, the NSF may consider (re)developing a funding program similar to VOSS, and encourage more involvement of social scientists who can carry out research on the five areas (and other important topics) under OAC, towards a strategic and collective effort of building smart CI in a sustainable way with widespread adoption and successful implementation. This program should be under OAC as the context of CI is unique. Social science conducted within the context of CI is critical for accurate applications.

References

[1] C. P. Lee, P. Dourish, and G. Mark, "The human infrastructure of cyberinfrastructure," CSCW '06: Proceedings of the Conference on Computer Supported Cooperative Work, pp. 483-492, New York: ACM Press, 2006.

[2] K. F. Kee, "Adoption and diffusion," In International encyclopedia of organizational communication, C. Scott and L. Lewis Eds. Hoboken, NJ: Wiley-Blackwell, 2017, pp. 41-54.

[3] B. Le, F. Escalera, K. Jitkajornwanich, and K. F. Kee, "External communication to diffuse science gateways and cyberinfrastructure as innovations for research with big data," Gateways Conference, San Diego, CA, 2019, available at: https://osf.io/mz8bv/

[4] D. Rock, and H. Grant, "Why diverse teams are smarter," Harvard Business Review, 2016, available at https://hbr.org/2016/11/why-diverse-teams-are-smarter

[5] S. L. Star, and G. C. Bowker, "How to infastructure," Handbook of new media: Social shaping and social consequences of ICTs, L. A. Lievrouw & S. M. Livingstone, Eds. Thansand Oaks, CA: Sage, 2006, pp. 230-245.

[6] D. Ribes, and T. A. Finholt, "The long now of technology infrastructure: Articulating tensions in development", J. of the Assoc for Information Systems, vol. 10, pp. 375-398, 2009.

Parallel Relational Algebra for Logical Inference at Scale

Sidharth Kumar, Thomas Gilray
University of Alabama at Birmingham
https://sidharthkumar.io — sid14@uab.edu
https://thomas.gilray.org/ — gilray@uab.edu

Abstract: Relational algebra (RA) comprises an important basis of operations, conspicuously sparse in the HPC literature. It can be used to implement a variety of algorithms in satisfiability and constraint solving [4], graph analytics [6], program analysis and verification [3], deductive databases [2], and machine learning [5]. Many of these applications are, at their heart, cases of logical inference; a basis of performant relational algebra is sufficient to power state-of-the-art forward reasoning engines for Datalog and related logic-programming languages. Declarative logic programming offers the promise of unifying specification and implementation, permitting programmers to focus on writing correct and maintainable code, and permitting its operational evaluation to be synthesized automatically. Despite its expressive power, relational algebra has not received the same attention in high-performance-computing research as more common primitives like stencil computations, floating-point operations, numerical integration, and sparse linear algebra. Furthermore, specific challenges in permitting fixed-point iteration, in addressing representation and communication among distributed portions of a relation, and in balancing inherently unbalanced relations, have previously thwarted successful scaling of relational algebra applications to HPC platforms. We are developing a set of efficient algorithms to effectively parallelize and scale key relational algebra primitives. We aim to develop foundational theory, practical implementations, and rigorous evaluations of our approach on three important application domains with the ultimate goal of enabling massively parallel distributed reasoning on cluster computers directed by expressive, ergonomic, rule-based languages.

Applications: To evaluate the scalability of our parallel RA infrastructure, we focus on three applications: graph mining **(E1)**, static program analysis **(E2)**, and deductive databases for physical simulation data **(E3)**. Figure 1 shows the overall pipeline we propose; the top shows our three experimental applications, sitting atop a platform for logical inference, implemented with relational algebra that the techniques under investigation **(T1-T3)** supports.

Typical graph computational algorithms are not suited for mining tasks that aim to discover complex structural patterns of a graph. Extracting such features such as paths, cliques, frequent subgraphs, etc, is straightforward to implement using relational algebra. We focus on two fundamental graph-mining tasks **(E1)**, transitive closure and clique computation. These applications are the most immediate uses of relational algebra; transitive closure is simply an iterated sequence of relational *join*, *projection*, and *union*, until a fixed point is reached. Second, we focus on static program analysis **(E2)**, a vital logical inference problem that exemplifies the power of our approach. Static analysis brings in substantial task- and data-parallelism, as well as evolution in balancing requirements across time. Finally, we focus on deductive database applications and the use of relational algebra as a combined storage and-inference system **(E3)**. Of particular interest to us is the possibility that logical inference can benefit traditional HPC problems for feature extraction in an automated way.

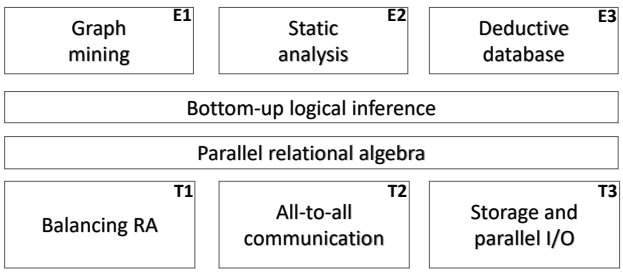


Figure 1: An overview of our pipeline: our application-foci are implemented as bottom-up logical inference, implemented as task- and data-parallel relational algebra, built using the specific techniques we investigate in our proposed work.

Techniques: One of the central challenges in parallel relational algebra is how to balance the various facets of the problem: balancing communication and computation, balancing across iterations of an inference task, balancing across relations, and balancing among task-level components. We have built on previous approaches by developing strategies that mitigate load-imbalance in a dynamic manner. Our own approach [1] uses a two-layered distributed hash-table to partition tuples over a fixed set of *buckets*, and, within each bucket, to a dynamic set of *subbuckets* which may vary across buckets. Each tuple is assigned to a bucket based on a hash of its key-column values, but within each bucket tuples are hashed on non-join-column values, assigning them to a local subbucket, then mapped to an MPI process. The first step in a join operation is an *intra-bucket communication* (Figure 2a) phase within each bucket so that every subbucket receives all tuples for the outer relation across all subbuckets (while the inner relation only needs tuples belonging to the local subbucket). Following this, a *local join* operation (with any necessary projection and renaming) is performed in every subbucket (Figure 2b), and, as output tuples may each belong to an arbitrary bucket in the output relation, an MPI *all-to-all* communication phase (Figure 2c) shuffles the output of all joins to their managing processes (preparing them for any subsequent iteration).

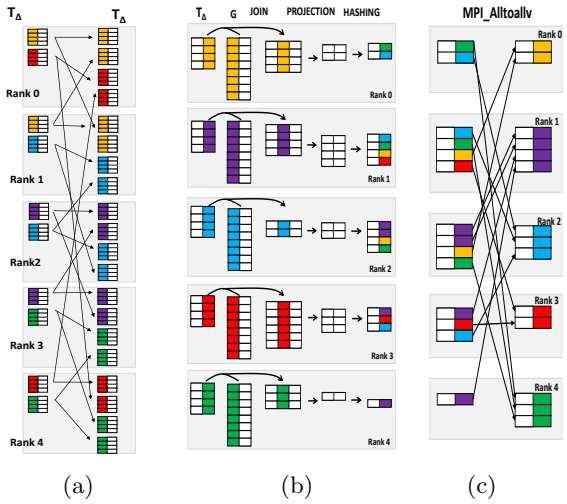


Figure 2: (a) Intra-bucket communication; each subbucket of T_Δ sends its data to all subbuckets of G. (b) Local, per-subbucket joins (including projection and re-hashing). (c) All to all communication.

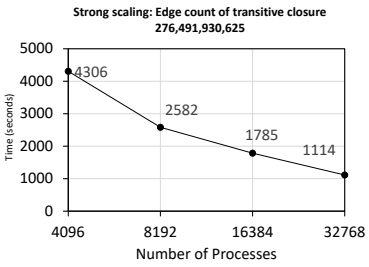


Figure 3: Strong scaling for TC computation of SuiteSparse graph `mc2depi`.

Results: The transitive closure (T) of an input graph (G) is iteratively extended by adding new paths discovered by a join operation until a fixed point is reached, and no new paths can be added to T. We performed strong scaling analysis for a graph with edge count 2,100,225 (mc2depi), varying the number of processes from 4,096 to 32,768. Figure 3 shows a preliminary strong-scaling study. We observed decent scaling to about 16k cores, producing a graph of over 276 billion paths. To the best of our knowledge, this is the largest such feature extraction now described in the literature. Given that this approach does not use any load balancing and does nothing to optimize a synchronous use of MPI's `All_to_allv` primitive, this scaling plot is a promising proof-of-concept that iterated joins on graphs derived from real applications will be scalable to 10k+ processes.

Conclusion and research direction: Effective declarative programming represents a long-standing dream of computing—exchanging code describing *how* to compute for code simply describing *what* to compute. Instead of requiring programmers to themselves balance the vying concerns of correctness, maintainability, and scalability in each task, declarative programming languages permit users to focus on the first two concerns, writing high-level specifications of *what* should be computed, while allowing the underlying implementation (i.e.,*how* the operational mechanics of the program work) to be extracted automatically. With this research we are making inroads by developing techniques for parallelizing relational algebra as a platform for declarative logical inference tasks. Our research is motivated by applications from the domain of graph analysis (feature extraction), static program analysis (reverse engineering, component verification, exploit generation, etc), and deductive databases (statistical and topological features).

References

[1] Sidharth Kumar and Thomas Gilray. Distributed relational algebra. In *International Conference on High Performance Computing, Data, and Analytics (In Submission)*, 2019.

[2] Mengchi Liu, Gillian Dobbie, and Tok Wang Ling. A logical foundation for deductive object-oriented databases. *ACM Transactions on Database Systems (TODS)*, 27(1):117–151, 2002.

[3] Bernhard Scholz, Herbert Jordan, Pavle Subotić, and Till Westmann. On fast large-scale program analysis in datalog. In *Proceedings of the 25th International Conference on Compiler Construction*, CC 2016, pages 196–206, New York, NY, USA, 2016. ACM.

[4] Emina Torlak and Daniel Jackson. Kodkod: A relational model finder. In *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, pages 632–647. Springer, 2007.

[5] Jurgen Anne Francios Marie Van Gael, Ralf Herbrich, and Thore Graepel. Machine learning using relational databases, January 29 2013. US Patent 8,364,612.

[6] Daniel Zinn, Haicheng Wu, Jin Wang, Molham Aref, and Sudhakar Yalamanchili. General-purpose join algorithms for large graph triangle listing on heterogeneous systems. In *Proceedings of the 9th Annual Workshop on General Purpose Processing Using Graphics Processing Unit*, GPGPU '16, pages 12–21, New York, NY, USA, 2016. ACM.

NSF Workshop on Developing a Roadmap towards the Next Generation of Smart Cyberinfrastructure

White Paper

Delivering AI and ML capable cyberinfrastructures as a Service

Ling Liu
School of CS, Georgia Institute of Technology

Artificial Intelligence (AI) and Machine Learning (ML) have penetrated every discipline of science and engineering. Many scientific discovery and engineering breakthrough today are empowered by AI and ML capabilities. NSF is in the unique position to develop the next generation of smart Cyberinfrastructure for supporting and facilitating AI-enabled scientific research discovery and engineering innovations in all sciences and engineering disciplines.

AI and ML capable hardware cyber-infrastructures.
As AI and ML capabilities rapidly evolve in science and engineering research and development efforts supported by NSF, it is vital to scale NSF supported cyber-infrastructure from experimentation to implementation, enabling and facilitating researchers and graduate students from different science and engineering disciplines to conduct field experiments using advanced AI-capable cyber infrastructures sponsored by NSF, successfully enabling universities and research institutes to achieve AI at scale. I below list three examples

(1) Huge Data Capable AI-ML model training infrastructures, including various scales of GPU clusters with industry strength huge model training capability.
(2) Large scale Federated Learning cyber infrastructure, including large scale AI-compute clusters, each with large number AI-capable computer nodes, enabling distributed federated training implementation and experimentation.
(3) A variety of EdgeSystems, enabling EdgeAI infrastructures for large scale implementation and experimentation of Edge AI model training, model prediction and active-learning.

Taking Federated Learning Infrastructures as an example. Traditionally, training ML models requires all data to be residing in the same trusted compute server. For huge data, this can be communication intensive in addition to privacy concerns and legislations such as General Data Protection Regulation (GDPR) and Health Insurance Portability and Accountability Act (HIPAA). Federated deep learning has emerged as an important alternative AI-capable cyberinfrastructure and distributed AI-training paradigm. Federated Learning (FL) has been fueled by a number of big data companies, represented by Google, Facebook, Amazon, Apple. In a typical federated deep learning system, each data owner (participant) maintains its own data locally and follows a federated learning protocol where only updates of the model training parameters are shared with the trusted parameter server (model training aggregator). Participants as workers are responsible for training the same model on different mini-batches of the huge data (compute intensive tasks). In each training iteration, each participant sends its local parameter updates to the parameter server, typically hosted

in the Cloud, which stores, aggregates and maintains a set of shared parameters. The parameter server shares the aggregated parameters with each of the participants in the federated learning system and each participant updates its local parameters in the subsequent iteration. This distributed training process iterates until the model training reaches the pre-defined convergence condition. In these FL scenarios, participants are heterogeneous compute nodes that communicate with the parameter server in either a client-server style or decentralized peer to peer style. We refer to this in-network memory computing paradigm a Cloud coordinated in-network memory computing.

Setting such federated learning cyberinfrastructure at scale is not only a huge challenge for all academic researchers and graduate students and it also requires large up front capital investments. By providing the NSF-sponsored smart cyberinfrastructures to facilitate federated learning, on one hand, it will fuel the research innovation and advancement in federated learning systems, algorithms and optimizations; and on the other hand, it will enable many scientific research labs to train on their local data collections and share only parameters, which can be significantly beneficial for hospitals, healthcare professionals, virus research laboratories, and so forth. Such federated learning empowerment will shorten the path and sparkle the lights towards new, transformative scientific discovery and engineering innovations.

AI and ML capable Software Cyberinfrastructures. The next generation smart cyberinfrastructure should also be equipped with AI and ML enabled software as a service platforms and tools. Such AI software agents can be instrumental in the event of rare virus spreading. For example, if there is a shortage of doctors, caregivers can remotely monitor at risk patients and robots can deliver medicines for humans to minimize infection spreading and help reduce reliance on doctors for routine readings and so forth.

In the digital transformation and big data age, AI is an important tool, which will make changes in the ways how scientists and engineers tackle problems in every area of our sciences and engineering and daily life. AI-capable software agents will make things easier for scientists and engineers to focus more closely on innovation and technology trends.

AI and ML Security and Privacy Cyberinfrastructures. AI models and ML models are vulnerable to adversarial perturbation attacks. A smart cyberinfrastructure should be capable of providing built-in secure AI workflows throughout the AI-training and AI prediction workflow life cycle, enabling data input guard, AI inference guards and AI output guards.

I have done research in the above areas, especially in federated learning infrastructures and secure AI/ML workflow system infrastructures. I am happy to share my experiences, expertise and lessons learned with the workshop attendees and contribute to discussions at the workshop and learn from other participants on their visions and experiences on the open challenges in building the next generation smart cyber-infrastructures at the workshop.

White Paper for NSF Smart Cyberinfrastructure (CI) Workshop

Washington DC Feb. 25-27, 2020

Naveen Sharma, Rochester Institute of Technology, Rochester, NY

We live in the era where the confluence of "big data" and software offers unprecedented opportunities to accelerate progress across all aspects of human endeavor. Big data is already enabling dramatic transformations in science, engineering, medicine, and public policy [1]. Consequently, we are witnessing an explosion of research and technology development initiatives that employ various data-driven methods under the broad umbrella of machine learning (ML) and data-driven artificial intelligence (AI). While ML/AI have gained much popularity and have shown notable successes in business for some time, fostered by industries such as e-commerce, marketing, and process optimization significant efforts are underway in its applications to natural sciences, where algorithms are used to stimulate scientific discovery. For example, advances in image processing, big data, data visualization, coupled with domain-specific computational models for climate science are enabling climatologists to model climate changes from data such as high spatial resolution (HSR) remote sensing images of sea ice (sea ice acts as both an indicator and an amplifier of climate change) [2]. Similarly, applications of ML/AI along with advances in information extraction and big data processing are enabling: biologists to gain insights into how living systems adapt; health scientists to devise targeted treatments and interventions to optimize health outcomes; education researchers to personalize pedagogy to optimize learning outcomes; social scientists to study why organizations, societies, and cultures succeed or fail; urban planners to design for optimal traffic flow; material scientists to develop new materials with properties not seen before. Progress in many areas of human endeavor is increasingly enabled by our ability to acquire, share, integrate, and analyze disparate types and modalities of data (i.e. big data) and new methods and tools for data integration, analysis, modeling, and interpretation. Holistically, seamlessly integrated "big data" and "software enabled capabilities" with networking, security resources, tools and services, and people skills collectively enables these new capabilities [3]. We advocate that to exercise their full potential future CIs ("Smart CI") will be highly programmable by the domain experts. Also, CIs enabling solutions to urban challenges will need to seamlessly integrate with the community and its citizenry.

Current ML and AI technologies do not provide easy ways for domain experts who are not ML/AI experts to develop applications. The fundamental goal of ML is to induce or synthesize programs that are able to learn from data. However, in current practice the programs are reduced to a model or set of models mapping inputs to outputs. Learning is an optimization process driven by an objective function of a predefined form [4]. Also, despite advances in "big data" the task of developing ML/AI applications is largely manual and labor-intensive. The real promise of ML/AI is rather limited to organizations possessing advance skills in computing, algorithms, and statistics. Thus, advancing the capability and capacity for ML/AI use in predictions and data-driven decision making in science, engineering, and public policy requires extensive support and involvement of skilled data scientists. Domain experts working with data scientists develop computational abstractions for relevant domains and associated methods and tools for domain data analysis, simulation, visualization, and sharing, and integration. Currently, multiple NSF-funded research efforts across wide array of domains are underway that focus on building domain-specific frameworks or platforms. Generally, these systems intend to provide good toolboxes and libraries for ML/AI based on existing techniques and domain-specific data in one place. Examples for such CI span across computational physics, climate modeling, cardiovascular simulations, material discovery, plasma physics, hydrologic modeling, and many more. In general, the focus is on building CIs comprising proven domain-specific models and abstractions along with integrated domain-data and views from various disparate sources (e.g. open

data sets). In most cases, ML models are the key component of these systems, but a typical solution involves multiple such models, along with significant levels of reasoning with the models' output and input. Current technologies do not make such techniques easy to use for domain experts who are not fluent in ML nor for ML experts who aim at testing ideas and models on real-world data in the context of the overall AI system. To realize full potential, we argue that the future CI systems will need to provide easy to use interface by being *highly programmable* – both to attract a larger user base as well as software evolution. Quickly, *variability* is the number of possible different evolutions of a system (CI) where as *programmability* is the capability of a system to change or to react to external stimuli (input) in order to alter its behavior [5]. A highly programmable CI system will raise the level of abstraction at which the user conceptualizes and develops ML/AI models. Lowering the technical barriers to access CI will enable easy and productive access to AI/ML tools for science as well as a large community of science stakeholders. On one hand, the complexity of ML/AI necessitates this capability for wider scale adoption; on the hand ML/AI techniques itself can help design this capability. Thus such a capability enables both *CI for ML/AI and ML/AI or CI* paradigms. A highly-programmable CI in a specific domain, e.g. hydrologic modeling or material discovery will most likely offer a domain-specific interface – i.e. a very high level and domain-specific programming language – conducive to that domain, it will enable synthesis of a set of programs implementing ML/AI models for the data presented. Users can pick and choose which models best fit their application. Such programmable interface will exercise CI in multitudes of innovative ways and far more than a canned interface.

By 2050 it is predicted that 66% of the global population will habitat in large- and mid-size cities. To date "smart city" solutions to issues of human development, while sometimes useful, do not get to the core of the urban issues, specifically at the community[a] and neighborhood levels. Having more high-quality data about the activities of citizens, households, and businesses in a community and having easy access to this data in a secured, timely and open fashion, can undoubtedly be of great help in designing service delivery systems or managing infrastructure efficiently. Community Based Participatory Research (CBPR) is widely used and has become a standard approach in a wide variety of domains, such as environmental and sustainability studies, criminology, community psychology, studies of race and gender, urban planning, community development and urban studies, migration studies and international development studies. "*CBPR emphasizes collaborative, equitable partnerships among researchers, stakeholders and community members throughout all phases of research. ... Communities are involved in decision-making throughout the research process, from developing research questions to disseminating research findings"* [6]. CBPR emphasizes the importance of identifying and validating the community's strengths and assets, avoiding an exclusive focus on problems. We argue that future CI systems focused on urban issues (or Urban CI) will need to seamlessly integrate with communities and neighborhoods empowering its citizenry. From driverless vehicles to software that runs subways systems to dynamic bus scheduling to smart grids, Urban CIs enable and help run a city more efficiently. However, over the long run, the parameters of these engineered systems and services, such as the quality of service in terms of use of time, comfort, economic cost, etc., must continuously adapt to societal issues. In fact, the Urban CIs will need to evolve as the community evolves. True integration with the community will enable community residents to use and contribute to Urban CI as part of evolution. In other words *Urban CIs shall be for the community and by the community!* Armed with data and access to analytics, communities can cut through ideological boundaries, focus on things that matter, and engage in conversations about challenges and opportunities. Such systems can engender processes with added benefit of creating a path of dialogue and inclusion to the urban poor and of responsive and knowledgeable government around practical issues for official organizations and private actors.

[a] A group of people living in the same defined area sharing the same basic values, organizations, interests, and sense of identity.

References

[1] Chen, H., Chiang, Roger H.L., and Storey Veda, C. “Business Intelligence and Analytics: From Big Data to Big Impact”

[2] Yang, Chaowei, Yu, Manzhu, Li, Yun, Hu, Fei, Jiang, Yongyao, Liu, Qian, Sha, Dexuan, Xu, Mengchao, and Gu, Juan. "Big Earth data analytics: a survey," Big Earth Data, v.3, 2019

[3] NSF’s Blueprint for a National Cyberinfrastructure Ecosystem “ Transforming Science Through Cyberinfrastructure”

[4] Parisa Kordjamshidia, Dan Roth, and Kristian Kerstingd “Declarative Learning-Based Programming as An Interface to AI Systems”

[5] Zenil, Hector “A Behavioural Foundation for Natural Computing and a Programmability Test”

[6] Collins, S. E. "Community-based participatory research (CBPR): Towards equitable involvement of community in psychology research community ," American Psychologist, vol. 73, pp. 884-898, 2018.

Convergence of Artificial Intelligence and High Performance Computing on NSF-supported Cyberinfrastructure

E. A. Huerta*†, Asad Khan*‡, Edward Davis*§, Colleen Bushell*, William D. Gropp*¶, Daniel S. Katz*¶‖**, Volodymyr Kindratenko*¶‖, Seid Koric*††, William T. C. Kramer*¶, Brendan McGinty*, Kenton McHenry*, Aaron Saxton*

*National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA

†Department of Astronomy, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA

‡Department of Physics, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA

§University of Queensland, St Lucia, QLD 4072 Australia

¶Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA

‖Department of Electrical & Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA

**School of Information Sciences, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA

††Department of Mechanical Science and Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA

***Abstract*—Significant investments to upgrade or construct large-scale scientific facilities demand commensurate investments in R&D to design algorithms and computing approaches to enable scientific and engineering breakthroughs in the big data era. The remarkable success of Artificial Intelligence (AI) algorithms to turn big-data challenges in industry and technology into transformational digital solutions that drive a multi-billion dollar industry, which play an ever increasing role shaping human social patterns, has promoted AI as the most sought after signal processing tool in big-data research. As AI continues to evolve into a computing tool endowed with statistical and mathematical rigor, and which encodes domain expertise to inform and inspire AI architectures and optimization algorithms, it has become apparent that single-GPU solutions for training, validation, and testing are no longer sufficient. This realization has been driving the confluence of AI and high performance computing (HPC) to reduce time-to-insight and to produce robust, reliable, trustworthy, and computationally efficient AI solutions. In this white paper, we present a summary of recent developments in this field, and discuss avenues to accelerate and streamline the use of HPC platforms to design accelerated AI algorithms.**

I. Introduction

The big data revolution disrupted the digital and computing landscape in the early 2010s [1]. Data torrents produced by corporations such as Google, Amazon, Facebook and YouTube, among others, presented a unique opportunity for innovation. Traditional signal processing tools and computing methodologies were inadequate to turn these big-data challenges into technological breakthroughs. A radical rethinking was urgently needed [2], [3].

Large Scale Visual Recognition Challenges [4] set the scene for the ongoing digital revolution. The quest for novel pattern recognition algorithms [5]–[7] that sift through large, high-quality data sets eventually led to a disruptive combination of deep learning and graphics processing units (GPUs) that enabled a rapid succession of advances in computer vision, speech recognition, natural language processing, and robotics, to mention just a few [8], [9]. These developments are currently powering the renaissance of AI, which is the engine of a multi-billion dollar industry.

Within just a few years, the emergence of high-quality data sets, e.g., `ImageNet` [10]; GPU-accelerated computing [11]; open source software platforms to design, train, validate and test AI models; improved AI architectures and novel techniques to enhance the performance of deep neural networks, such as robust optimizers and regularization techniques, led to the rapid development of AI tools that significantly outperform other signal processing tools on many tasks. These developments have been astonishing to witness. Data-driven discovery is now also informing and stirring the design of exascale cyberinfrastructure, in which HPC and data have become a single entity, namely HPCD [2], [12].

II. Convergence of AI and HPC

The convergence of AI and HPC is being pursued in earnest across the HPC ecosystem. Recent accomplishments of this program have been reported in plasma physics [13], cosmology [14], gravitational wave astrophysics [15], multi-messenger astrophysics [16], materials science [17], data management [18], [19] of unstructured datasets, and genetic data [20], among others.

These achievements share a common thread, namely, the algorithms developed to accelerate the training of AI models in HPC platforms have a strong experimental component. To date, there is no rigorous framework to constrain the ideal set of hyper-parameters that ensures rapid convergence and optimal performance of AI models as the number of GPU nodes is increased to accelerate the training stage.

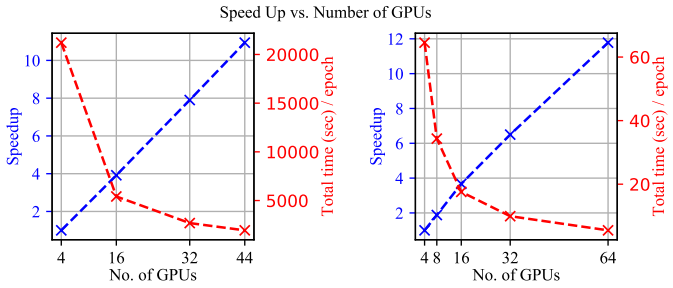


Fig. 1. Left panel: the training of an AI model (which characterizes the signal manifold of spinning binary black hole mergers) is reduced, achieving state-of-the-art performance, from one week to 17 hours by distributing the training workload up to up to 44 V100 GPUs using `Horovod` on the HAL cluster. Right panel: as the left panel, but now the training is reduced from 5 hours to 90 seconds using the entire HAL cluster. This AI model is used to classify and label galaxy images observed by two electromagnetic surveys.

In the context of NSF-supported infrastructure, we present two sample cases of AI and HPC convergence using the Hardware-Accelerated Learning (HAL) cluster [21] at NCSA.

The HAL cluster has 64 NVIDIA V100 GPUs distributed evenly across 16 nodes, and connected by NVLink 2.0 [21] inside the nodes and EDR InfiniBand across the nodes. Using this system, Figure 1 presents two science drivers: (1) an AI model to characterize the signal manifold of binary black mergers that is trained with time-series signals that describe gravitational wave signals [22]; (2) an AI model that classifies galaxy images collected by the Sloan Digital Sky Survey (SDSS), and automatically labels images collected by the Dark Energy Survey (DES) [14]. Using a single V100 GPU, these two models take on average one week and five hours to train, respectively. However, when the training of these models is distributed across HAL, one may fully train them achieving state-of-the-art performance within 17 hours and 90 seconds, respectively.

These examples clearly underscore the importance of coupling AI with HPC, i.e., accelerating the training stage enables: (1) the exploration of domain-inspired architectures and optimization schemes that are critical for the design of rigorous, trustworthy and interpretable AI solutions; (ii) the use of larger training data sets to boost the accuracy and reliability of AI models.

III. Software and Hardware Challenges

While open source software platforms have played a key role in the swift evolution of AI, they present a number of challenges when used in HPC platforms. This is because open source software platforms such as `TensorFlow` [23] and `PyTorch` [24] are updated at a much faster pace than libraries deployed cluster-wide on HPC platforms. Furthermore, producing AI models usually requires a unique set of package dependencies. Therefore, the traditional use of modules has limited effectiveness since software dependencies change between projects and sometimes evolve even during a single project. Common solutions to give users more fine-grained control over software environments include containerization (e.g., `Singularity` [25] or `Kubernetes` [26]), and virtual environments (e.g., `Anaconda` [27], which is extensively used by deep learning practitioners). We provide below a number of recommendations to streamline the use of HPC resources for AI research:

1) Provide up-to-date documentation and tutorials to set up containers and virtual environments, and adequate help desk support to enable smooth, fast-paced project life-cycles.
2) Maintain a versatile, up-to-date base container image, and base virtual environment that users can easily clone and modify for their specific needs.
3) Distributed training software stacks such as `TensorFlow` depend on distributed training software stacks (e.g., `Horovod` [28]), which in turn depend on system architecture and specific versions of `MPI` installed by the root administrator. It is important to have clear up-to-date documentation on system architecture and `MPI` versions installed, and clear instructions on how to install/update distributed training software packages like `Horovod` into the user's container/virtual environment.

In addition to these considerations, the AI model architecture, data set, and training optimizer prevent a seamless use of distributed training. Stochastic gradient decent (SGD) and its variants are the workhorse optimizer for AI training. The common way to parallelize training is to use "mini-batches" with SGD. In principle, a larger mini-batch may naively utilize more GPUs (or CPUs). Training time to solution will often scale linearly with small batch size. Figure 1 shows good generalization at 64 GPUs, which amounts to a global batch size of 128 samples. However, it is known that as data sets and number of features grow, naively scaling number of GPUs, and subsequently batch size, will often take more epochs to achieve an acceptable validation error. The state-of-the art in AI training at scale was reported in [29], who trained `ResNet-50` using a batch size of 64k samples, run across 2048 Tesla P40s. While achieving this level of scaling required a lot of experimental work, this benchmark, and others [30], indicate that scaling AI models to larger data and feature sets is indeed possible. However, it requires a considerable amount of human effort to tune the model and training pipeline. A mixture of fast human model development cycle mixed with automated hyperparameter tuning is a candidate solution to tackle this problem.

IV. Cloud Computing and HPC

Cloud computing and containerization became popular for developing customer facing web apps. It allowed a DevOps team to keep strict control of the customer facing software, while new features and bug fixes were designed, developed, and tested in an environment that "looked the same" as a live one. Depending on the business cycle, companies could dynamically scale their infrastructure with virtually no overhead

of purchasing hardware, and then relinquish it when it was no longer needed.

HPC would do well to adopt a DevOps cycle like the ones seen in startup culture. However HPC has some unique challenges that make this difficult. 1) Data storage separated from compute in the form of a shared file system and an instance on maintaining a traditional tree like file system. Cloud computing delivers a unit of compute and storage in tandem as a single instance and isolates distinct resources. A developer using cloud resources treats a compute instance as only the host for their code and must explicitly choose how to move large volumes of data on and off. This is usually done by allocating a specialized cloud instance of a data store (e.g., SQL databases). Improved cloud solutions provide `Kubernetes` (and other cluster manager) recipes to allocate a skeleton of these resources, but it is still up to the developers to choose exactly how data are moved between the resources and to code the specific functions of their app. 2) HPC is a shared resource. That is, many users with different projects see the same file system and compute resource. Each developer must wait their turn to see their code run. In cloud computing, a resource belongs and is billed to the developer on demand. When the resource is released, all of its state-full properties get reset. 3) HPC is very concerned with the compute resources interconnect. To have high bandwidth and low latency between cloud compute instances, one pays a premium.

In the case of distributed training, one needs to ascertain whether the cloud or HPC platforms provide an adequate solution. On-demand, high throughput or cloudbursting of single-node applications are ideally suited for the cloud. For instance, in the case of genetic data analysis, the `KnowEng` platform [20] is implemented as a web application where the compute cluster is managed by `Kubernetes`, and provides an example of a workflow that can be expanded to include methods for intuitively managing library compatibility and cloud bursting. This cloud-based solution includes: (1) the ability to access disparate data; (2) set parameters for complex AI experiments effortlessly; (3) deploy computation in a cloud environment; (4) engage with sophisticated visualization tools to evaluate data and study results; and (5) save results and access parameter settings of prior runs.

However, large distributed training workloads, that run for many hours or days will continue to excel on a high-end HPC environment. For instance, the typical utilization of the HAL cluster at NCSA, which tends to be well above 70%, would require a monthly investment of around $100k in comparable cloud compute resources; this is far higher than the amortized cost of the HAL cluster and its support.

V. Industry Applications

The confluence of AI and HPC is a booming enterprise in the private sector. NCSA is spearheading its application to support industry partners from the agriculture, healthcare, energy, and financial, sectors to stay competitive on the global market by analyzing bigger and more complex data to uncover hidden patterns, reveal market and cash flow trends, and identify customer preferences. The confluence of modeling and simulation and AI is another area of growing interest among manufacturing and life science partners, promising to significantly accelerate many extremely difficult and computationally expensive methods and workflows in model-based design and analysis [31]–[33].

Cross-pollination in AI research between academia and industry will continue to inform these activities, making an optimal use of HPC and cloud resources, to design and deploy solutions that transform AI innovation into tangible societal as well as business benefits.

VI. Conclusion

The convergence of AI and HPC is strongly poised to fully exploit the potential of AI in science, engineering and industry. Realizing this goal demands a concerted effort between AI practitioners, HPC and domain experts. It is essential to design and deploy commodity software across HPC platforms to facilitate a seamless use of state-of-the-art open source software platforms for AI research. It is urgent to go beyond experimental approaches that lack generality to optimally use oversubscribed NSF resources. An initial step in this direction includes making open source existing solutions that scale well while exhibiting good generalization in mid-scale clusters.

Acknowledgment

EAH, AK, DSK, and VK gratefully acknowledge National Science Foundation (NSF) awards OAC-1931561. EAH and VK also acknowledge NSF award OAC-1934757. This work utilized XSEDE resources through the NSF award TG-PHY160053, and the NSF's Major Research Instrumentation program, award OAC-1725729, as well as the University of Illinois at Urbana-Champaign.

References

[1] M. Asch, T. Moore, R. Badia, M. Beck, P. Beckman, T. Bidot, F. Bodin, F. Cappello, A. Choudhary, B. de Supinski, E. Deelman, J. Dongarra, A. Dubey, G. Fox, H. Fu, S. Girona, W. Gropp, M. Heroux, Y. Ishikawa, K. Keahey, D. Keyes, W. Kramer, J.-F. Lavignon, Y. Lu, S. Matsuoka, B. Mohr, D. Reed, S. Requena, J. Saltz, T. Schulthess, R. Stevens, M. Swany, A. Szalay, W. Tang, G. Varoquaux, J.-P. Vilotte, R. Wisniewski, Z. Xu, and I. Zacharov, "Big data and extreme-scale computing: Pathways to convergence-toward a shaping strategy for a future software and data ecosystem for scientific inquiry," *The International Journal of High Performance Computing Applications*, vol. 32, no. 4, pp. 435–479, 2018.

[2] National Academies of Sciences, Engineering, and Medicine, *Opportunities from the Integration of Simulation Science and Data Science: Proceedings of a Workshop*. Washington, DC: The National Academies Press, 2018.

[3] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[4] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[5] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov 1998.

[6] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[7] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989. [Online]. Available: https://doi.org/10.1162/neco.1989.1.4.541

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778.

[9] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. The MIT Press, 2016.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[11] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," *NIPS*, 2012.

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

[13] A. Svyatkovskiy, J. Kates-Harbeck, and W. Tang, "Training distributed deep recurrent neural networks with mixed precision on gpu clusters," in *Proceedings of the Machine Learning on HPC Environments*, ser. MLHPC'17. New York, NY, USA: Association for Computing Machinery, 2017. [Online]. Available: https://doi.org/10.1145/3146347.3146358

[14] A. Khan, E. A. Huerta, S. Wang, R. Gruendl, E. Jennings, and H. Zheng, "Deep learning at scale for the construction of galaxy catalogs in the Dark Energy Survey," *Physics Letters B*, vol. 795, pp. 248–258, Aug 2019.

[15] H. Shen, E. A. Huerta, and Z. Zhao, "Deep Learning at Scale for Gravitational Wave Parameter Estimation of Binary Black Hole Mergers," *arXiv e-prints*, p. arXiv:1903.01998, Mar 2019.

[16] E. A. Huerta *et al.*, "Enabling real-time multi-messenger astrophysics discoveries with deep learning," *Nature Rev. Phys.*, vol. 1, pp. 600–608, 2019.

[17] L. Ward, B. Blaiszik, I. Foster, R. S. Assary, B. Narayanan, and L. Curtiss, "Machine learning prediction of accurate atomization energies of organic molecules from low-fidelity quantum chemical calculations," *MRS Communications*, vol. 9, no. 3, p. 891–899, 2019.

[18] L. Marini, R. Gutierrez-Polo, R. Kooper, S. Satheesan, M. Burrnette, T. Nichoson, O. M. Zhao, Y., J. Lee, and K. McHenry, "Clowder: Open source data management for long tail data," *PEARC*, 2018.

[19] S. Padhy, J. Alameda, E. Black, D. M. K. P. Diesendruck, L., R. Kooper, J. Lee, R. Liu, R. Marciano, L. Marini, D. Mattson, B. Minsker, C. Navarro, M. Slavenas, W. Sullivan, J. Votava, and K. McHenry, "Brown dog: Leveraging everything towards autocuration," *IEEE BigData*, 2015.

[20] C. Blatti, A. Emad, M. J. Berry, L. Gatzke, M. Epstein, D. Lanier, P. Rizal, J. Ge, X. Liao, O. Sobh, M. Lambert, C. S. Post, J. Xiao, P. Groves, A. T. Epstein, X. Chen, S. Srinivasan, E. Lehnert, K. R. Kalari, L. Wang, R. M. Weinshilboum, J. S. Song, C. V. Jongeneel, J. Han, U. Ravaioli, N. Sobh, C. B. Bushell, and S. Sinha, "Knowledge-guided analysis of 'omics' data using the knoweng cloud platform," *bioRxiv*, 2019. [Online]. Available: https://www.biorxiv.org/content/early/2019/05/19/642124

[21] NCSA, "HAL Cluster," https://wiki.ncsa.illinois.edu/display/ISL20/HAL+cluster.

[22] A. Das, A. Khan, and E. A. Huerta, "The signal manifold of spinning binary black hole mergers. A Deep Learning Perspective," *In Preparation*.

[23] M. Abadi, A. Agarwal *et al.*, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," *ArXiv e-prints*, Mar. 2016.

[24] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS-W*, 2017.

[25] G. M. Kurtzer, "Singularity 2.1.2 - Linux application and environment containers for science," Aug. 2016. [Online]. Available: https://doi.org/10.5281/zenodo.60736

[26] Kubernetes, https://kubernetes.io/.

[27] Anaconda, https://www.anaconda.com/.

[28] A. Sergeev and M. Del Balso, "Horovod: fast and easy distributed deep learning in TensorFlow," *ArXiv e-prints*, Feb. 2018.

[29] X. Jia, S. Song, W. He, Y. Wang, H. Rong, F. Zhou, L. Xie, Z. Guo, Y. Yang, L. Yu, T. Chen, G. Hu, S. Shi, and X. Chu, "Highly scalable deep learning training system with mixed-precision: Training imagenet in four minutes," 07 2018.

[30] Y. You, Z. Zhang, C.-J. Hsieh, J. Demmel, and K. Keutzer, "Imagenet training in minutes," 2018. [Online]. Available: https://doi.org/10.1145/3225058.3225069

[31] D. W. Abueidda, S. Koric, and N. A. Sobh, "Machine learning accelerated topology optimization of nonlinear structures," *arXiv e-prints*, p. arXiv:2002.01896, Jan 2020.

[32] S. Luo, J. Cui, M. Vellakal, J. Liu, E. Jiang, S. Koric, and V. Kindratenko, "Review and Examination of Input Feature Preparation Methods and Machine Learning Models for Turbulence Modeling," *arXiv e-prints*, p. arXiv:2001.05485, Jan 2020.

[33] S. G. Rosofsky and E. A. Huerta, "Artificial neural network subgrid models of 2-D compressible magnetohydrodynamic turbulence," *arXiv e-prints*, p. arXiv:1912.11073, Dec 2019.

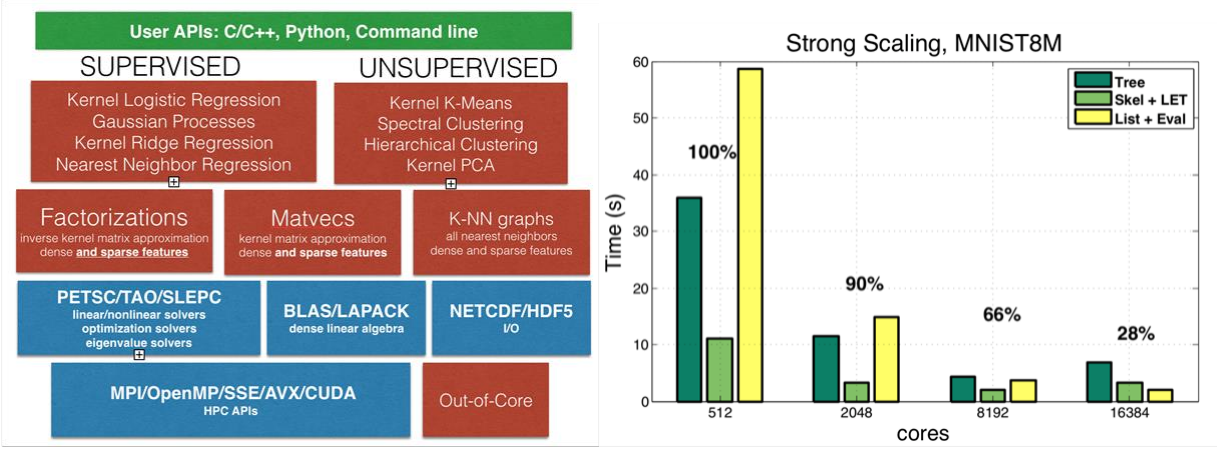
HPC Algorithms for Scalable Machine Learning and Data Analytics

George Biros
Parallel Algorithms for Data Analytics and Simulation Group
Oden Institute, The University of Texas at Austin

Increasingly, scientists and engineers are incorporating data analytics tasks in their workflows. Examples, of such tasks include uncertainty quantification, inverse problems, design and control, classification, anomaly detection, rare-event detection, data-assimilation, model reduction, pattern discovery, graph analytics, compression, visualization and dimensionality reduction to name a few. Many high-fidelity simulations in science and engineering require High Performance Computing (HPC) resources. Therefore, to increase the coherence between simulation and data analytics we need software that scales to large datasets and can support current and future HPC architectures.

Many groups in industry, government labs, and academia are working towards such HPC machine learning and data analytics software. However, the need to process in-situ large scientific and engineering datasets is not fully met with current software, and sometimes significant down-sampling is required in order to use existing tools. Also, popular ML libraries like Torch and TensorFlow are designed around industrial applications (natural language processing, audio, images, and video) and not scientific applications that have more unstructured data.

In my group, we have been focusing on algorithms and software for HPC machine learning algorithms with a focus on nearest neighbor graphs and kernel methods. (See figure below.)



***Left**: Work on HPC algorithms for machine learning in the PADAS group at UT. Red boxes indicate components developed in our group and support supervised and unsupervised learning. Different modules for kernel approximations, randomized nearest-neighbors, factorization of kernel matrices, kernel PCA, and toward supervised and unsupervised learning. **Right:** Scalability of kernel regression on a dataset with 8 million points in 768 dimensions. The run took place on 16,384 of TACC's Stampede I system.*

This research is funded by NSF awards CCF-1817048 and CCF-1725743.

Examples of kernel methods include support vector machines, logistic regression, Gaussian process approximation, density estimation, and clustering. These methods are in turn used in numerous applications in signal analysis, text analysis, and bioinformatics, and science and engineering. A key bottleneck in kernel methods is that they scale quadratically with the problem size. The problem size is typically defined as the number of examples (or training points)

presented to the machine learning algorithm. Since training points can number in the millions, various approximation schemes are necessary for scalability to large-scale problems. Software for kernel methods has to meet several conflicting goals: it should use of state-of-the art algorithms (which may be complicated and difficult to parallelize), achieve both shared memory and distributed memory scalability so that large-scale problems can be solved, be robust and easy to use (no excessive parameter tuning), achieve high-performance so that the underling hardware resources are efficiently utilized, and offer a sufficiently general suite of algorithms. Most of the existing machine packages either avoid kernel methods due to their complexity (related to the need to approximate dense matrices), or when they include kernel methods they lack accuracy (due to crude approximations) and scalability---particularly on distributed memory architectures. We have developed several C++ libraries that provide scalable implementations for a large class of problems. With support by NSF we have developed scalable software for kernel methods: ASKIT (library for kernel density estimation and kernel regression), RKDT (nearest neighbor algorithms), GOFMM (geometry oblivious hierarchical matrices that can be used to ``invert'' kernel matrices), and KLR (a library for kernel logistic regression). These packages can be found in https://padas.oden.utexas.edu. These algorithms have been ported on several XSEDE platforms at TACC and have been scaled to billions of points. Scalability is achieved by using a combination of MPI, OpenMP and GPU acceleration.

Outstanding challenges / requirements

- Performance portability on upcoming heterogeneous architectures.
- Utilizing variable precision algorithms.
- Support of development and deployment of HPDA algorithms and software that supports a large variety of structured and unstructured data.
- Support for easy to use interfaces, e.g., using Python or other high-level languages.
- Simpler interfaces to leadership resources.
- Long term publicly visible data repositories.
- Reproducibility of complex workflows with dependencies to ever changing libraries and large data.

Strengthening the Adoption of AI in Research and Cyberinfrastructure

Paola A. Buitrago
paola@psc.edu
Pittsburgh Supercomputing Center
Carnegie Mellon University
Pittsburgh, Pennsylvania

Nicholas A. Nystrom
nystrom@psc.edu
Pittsburgh Supercomputing Center
Carnegie Mellon University
Pittsburgh, Pennsylvania

INTRODUCTION

Artificial intelligence (AI) and machine learning (ML) have vast potential to advance research, and their increasing effectiveness will both shape and enhance cyberinfrastructure (CI) to enable the breakthroughs of the future. The rapidly expanding array of algorithms and technologies is a transformation with potential exceeding that of other architectural advances over the past few decades -– for example, simulations are being accelerated by factors of up to two billion [3], vastly outpacing what has been possible through Dennard scaling – and it is also the greatest fundamental shift in how researchers engage in computational science and engineering.

We face tremendous opportunities and exciting challenges in helping the research community realize the benefits of AI and ML. Hardware resources must continue to integrate high performance computing (HPC) and scalable AI resources effectively into heterogeneous systems, including a range of special-purpose AI architectures that will be superior for specific workloads, and provide user-friendly software environments to make the resources accessible to domain specialists (i.e., not only traditional HPC users). User training of many kinds is, and will continue to be, absolutely essential. Equally exciting is the prospect of applying AI and ML to the operation of cyberinfrastructure, or *Smart CI*, which has begun yet has considerable untapped potential. The increasing complexity of heterogeneous CI and the extremely diverse workloads that are executed on it make Smart CI vitally important to maximize the resources' potential and users' productivity.

1 OPEN RESEARCH CHALLENGES AND PROMISING RESEARCH DIRECTIONS

Fundamental AI research is flourishing [6] and beyond the scope of this white paper. Instead, we focus on three specific, promising areas of research that are of specific relevance to NSF-supported cyberinfrastructure: 1) scalable AI for scientific data; 2) AI-accelerated simulation; and 3) AI for system operations, *AIOps*. For all three, there is an opportunity to identify use cases that go beyond commercial ones that will inform selection of appropriate hardware technologies for production research platforms.

Scalable AI for Scientific Data: Research is needed to provide more robust tools for data acquisition and preparation, to increase the scalability of deep learning training (which will directly benefit use of generative adversarial networks (GANs) and deep reinforcement learning, methods that are being used to address challenges involving limited or sensitive data), to work with multimodal data, and in many other areas. Applications are analyses of very large datasets from instruments or simulations, identifying rare events, combining different kinds of data to allow answering new questions, facilitating the discover and reuse of scientific data, and helping to find relevant insights from scientific literature.

AI-Accelerated Simulation: Research is needed to improve the development of surrogate models for accelerating and potentially improving the accuracy of simulations, including scalable online training (i.e., incorporating new data into a model on an ongoing basis) for use with ensembles of simulations and data assimilation. Published results already demonstrate up to nine orders of magnitude acceleration [3] with no loss of accuracy [7], and impact on other applications is probable given sufficient human resources and user training. Applications are widespread, including both nontraditional fields (genomics, protein signaling pathways, 3D image processing, etc.) and traditional HPC (weather, chemistry, astrophysics, engineering, etc.).

AIOps: There are many promising directions, limited primarily by human resources, to improve CI through AIOps. Applying AIOps to system instrumentation data can improve system operation, reliability, and performance by identifying degrading components for proactive repair, spotting issues that negatively impact system performance, and optimizing queue structure. AIOps can improve users' experience, particularly for nontraditional users who are unaccustomed to advanced CI environments, as is being addressed by PSC's internal project *Calima*, led by Buitrago, for large systems such as *Bridges* [5], *Bridges-AI* [2], and *Bridges-2*. AIOps can also help with resource requests by helping to match advanced, heterogeneous computational resources to project requirements, helping to match reviewers to resource proposals, and then assessing whether the expected resource utilization is achieved.

2 TESTBEDS

Innovative hardware designed for specific aspects of AI and ML is being developed at an unprecedented pace. For example, emerging architectures accelerate deep learning training by orders of magnitude, deliver a petaflop/s on a chip for streaming inference, and maximize power efficiency for IoT sensor arrays.

It is essential for the research community to have access to a testbed (or testbeds) of such technologies to gain timely experience and identify those technologies that would benefit science and engineering. That knowledge would then inform larger-scale deployments of the technologies that prove to be most valuable. In many cases, architectural and AI expertise is needed to fully exploit new hardware technologies: simply providing users with access to newly emerging hardware is not sufficient. This kind of testbed is exactly what PSC is doing through its NSF-supported *Open Compass* [1] project,

which is evaluating the performance of deep learning networks relevant to research (which differ significantly from industry-motivated benchmarks such as MLPerf and GLUE) on new AI technologies, leveraging a frequently-refreshed AI technology testbed at PSC. Sustaining and expanding such effort is vital for effectively bringing new AI technologies to science and engineering research.

3 INNOVATIONS

Wide-ranging innovations will allow CI for AI-enabled science to reach its full potential. These innovations range across all aspects of resource provisioning, operations, and user engagement. Some examples are as follows.

AIOps can improve system reliability, availability, and performance. Early work has already been done to applying machine learning to system instrumentation data, for example, to identify and proactively repair components that are degrading.

The research community needs specialized, advanced hardware that delivers great scalability for specific types of problems. (This is in addition to testbeds of frequently-refreshed hardware as described in §2.) For example, individual training runs can take hours to weeks even on today's most advanced GPUs, and fully training a model can take hundreds or thousands of runs across different network architectures and sets of hyperparameters. New kinds of hardware that are architected specifically to accelerate deep learning training can accelerate those runs by orders of magnitude, potentially transforming researchers' ability to produce higher-accuracy models. Other kinds of specialized hardware are being developed for inferencing, streaming data, and edge devices (which can stream data back to large CI resources).

The research community also needs more flexible ways to build customized software environments. Tools such as Anaconda work for some software stacks, but not for all. Containers are an increasingly popular alternative, and container technologies such as Singularity [4] are widely adopted in HPC. However, many users lack local resources to build their own containers, and for some purposes (e.g., to incorporate system-specific libraries) it would be extremely helpful to build containers on the target platform. For that, it would be helpful to deploy "sandbox" environments which provide isolated, secure spaces in which users can have the necessary privilege levels to build or refine the containers they need.

The complexity of the AI landscape also introduces challenges in applying for, and granting, research requests. Innovations are needed in the process, guidelines, and tools to properly match project and compute needs to the right resources. Improved methods are needed for handling research proposals to accommodate the different, non-traditional nature of AI/ML and data science research. There is also an opportunity to use AIOps to improve the proposal review process, specifically by applying natural language understanding to infer the actual emphasis of each proposal, to assist in recruiting and assigning the most appropriate reviewers.

4 TRAINING

User training in scalable AI for NSF-supported CI is greatly needed and in very high demand. PSC leads the XSEDE HPC Monthly Workshop series [8], which rotates through topics of MPI, OpenMP, OpenACC, and Big Data & AI. To date, these workshops have hosted 11,921 participants. The most popular workshops, which are now offered every other month, are on Big Data & AI, for which 7,158 individuals from 83 different institutions participated in 19 events. Up to 26 institutions and 612 individuals have participated per event, using PSC's Wide Area Classroom format, and each event features extensive hands-on exercises.

The XSEDE workshops have proven to be extremely successful and valuable. A valuable next step would be to develop more advanced modules focusing on AI to foster workforce development and more effective use of advanced CI. Specifically, users need guidance on the following:

- *Data acquisition and preparation*, including tools for orchestrating training (e.g., hyperparameter optimization) and measurement of bias and correcting for it.
- *Scalable AI*, or using many AI accelerators together, including distributed across many nodes and the use of advanced hardware, to reduce the time needed to train models, thereby facilitating higher accuracy and testing more ideas.
- *Measuring and improving performance*, including profiling, I/O optimization, and maximizing the efficiency of task and data parallelism.
- *Choosing optimal resources and estimating requirements*, including how to determine which hardware technologies and software frameworks would best fit specific tasks and how to estimate computational requirements.
- *Advanced topics*, including explainability, developing and validating surrogate models, multimodal data, reinforcement learning, and domain-specific networks and data types.

5 TOOLS AND TECHNIQUES

Usability and effective tools are vital for CI- and AI-enabled science and engineering. Users increasingly come from laptop and cloud backgrounds where providers have prioritized usability to gain market share. CI providers must place the user first and deliver the following:

- *Highly usable systems* on par with cloud providers.
- *Intuitive, interactive monitoring tools* for both service providers and end users.
- *Tools to assist with data acquisition and preparation*, such as cleaning and labeling.
- *Flexible data handling tools* to tackle the complex workflows that occur when using AI for in a new field or new domain challenge.

SUMMARY

The resurgence of innovation in computer architecture and the high scientific impact of artificial intelligence and machine learning present unique opportunities for deploying transformative cyberinfrastructure platforms, enhancing their operation through use of AIOps, engaging in research on AI & ML issues specific to computational science, training the computational science community, and strengthening the software ecosystem with emphases on usability and data. For each of these areas, the Pittsburgh Supercomputing Center has initiatives already underway, yet significantly more effort is needed.

REFERENCES

[1] Paola A. Buitrago and Nicholas A. Nystrom. 2019. Open Compass: Accelerating the Adoption of AI in Open Research. In *Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (Learning)* (Chicago, IL, USA) *(PEARC '19)*. Association for Computing Machinery, New York, NY, USA, Article Article 72, 9 pages. https://doi.org/10.1145/3332186.3332253

[2] Paola A. Buitrago, Nicholas A. Nystrom, Rajarsi Gupta, and Joel Saltz. 2020. Delivering Scalable Deep Learning to Research with Bridges-AI. In *High Performance Computing: 6th Latin American Conference, CARLA 2019*, J. L. Crespo-Mariño and E. Meneses-Rojas (Eds.). Communications in Computer and Information Science, Vol. 1087. Springer, Basel, Switzerland.

[3] M. F. Kasim, D. Watson-Parris, L. Deaconu, S. Oliver, P. Hatfield, D. H. Froula, G. Gregori, M. Jarvis, S. Khatiwala, J. Korenaga, J. Topp-Mugglestone, E. Viezzer, and S. M. Vinko. 2020. Up to two billion times acceleration of scientific simulations with deep neural architecture search. arXiv:stat.ML/2001.08055

[4] Gregory M. Kurtzer, Vanessa Sochat, and Michael W. Bauer. 2017. Singularity: Scientific containers for mobility of compute. *PLoS ONE* 12, 5 (2017), e0177459. https://doi.org/10.1371/journal.pone.0177459

[5] Nicholas A Nystrom, Paola A Buitrago, and Philip D Blood. 2019. Bridges: Converging HPC, AI, and Big Data for Enabling Discovery. In *Contemporary High Performance Computing: From Petascale toward Exascale, Volume Three*, Jeffrey S. Vetter (Ed.). CRC Press, Boca Raton, FL.

[6] Raymond Perrault, Yoav Shoham, Erik Brynjolfsson, Jack Clark, John Etchemendy, Barbara Grosz, Terah Lyons, James Manyika, Saurabh Mishra, and Juan Carlos Niebles. 2019. *The AI Index 2019 Annual Report*. Technical Report. AI Index Steering Committee, Human-Centered AI Institute, Stanford University.

[7] Justin S Smith, Benjamin T Nebgen, Roman Zubatyuk, Nicholas Lubbers, Christian Devereux, Kipton Barros, Sergei Tretiak, Olexandr Isayev, and Adrian E Roitberg. 2019. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nature Communications* 10, 1 (2019), 2903. https://doi.org/10.1038/s41467-019-10827-4

[8] John Urbanic and Thomas Maiden. 2018. Evaluating the Wide Area Classroom after 10,500 HPC Students. In *2018 IEEE/ACM Workshop on Education for High-Performance Computing (EduHPC)*. IEEE, New York, NY, 51–60.

Developing a Roadmap towards the Next Generation of Smart Cyberinfrastructure
February 25-27, 2020, Hyatt Regency Crystal City, VA,

Intelligent Data Analytics Environment (IDAE) for High Performance Cyberinfrastructure

Salim Hariri
NSF I/UCRC Center for Cloud and Autonomic Computing
University of Arizona
Nsfcac.arizona.edu
hariri@email.arizona.edu

The current machine learning algorithms are mainly developed to run on sequential platforms. With the current exponential growth in data, and the large scale of ML applications, it is becoming critically important to reduce the execution time and improve the scalability of the ML algorithms. The main goal of this whitepaper is to highlight the main research challenges that must be addressed to allow high performance cyberinfrastructure to exploit the emerging big data programming paradigms (MapReduce), artificial intelligence, machine learning algorithms, and high performance platforms (parallel, distributed and clusters of GPUs).

Research Challenges

- ***How to handle large-scale dynamic and heterogeneous data streams?***
  - ***The main research problem is here is how to detect accurately the changes in data streams***
- ***How to develop an intelligent recommender system that tells the user the best Ml algorithm to use and how it should be configured?***
- ***How to apply parallel/distributed algorithms and big data programming tools to speedup ML computations?***
- ***How do you validate and benchmark the proposed approach on a wide range of scientific and engineering applications.***

In this white paper, we will discuss ongoing research activities at the University of Arizona to address these challenges. Figure 1 shows an environment to develop an Intelligent Data Analytics Environment (IDEA) and Figure 2 shows how to detect changes in the data streams being analyzed. Figure 3 shows an approach to adopt the Ml algorithm so it can model the recent changes in data streams. The first three figures can potentially address the first two research challenges.
Figure 4 shows a High Performance Machine Learning Framework (HPMLF) that can overcome the last two research challenges. The HPMLF will leverage Big Data analytics tools, MapRedue, parallel/distributed algorithms and high performance platforms (Parallel/distributed systems and GPU cluster).

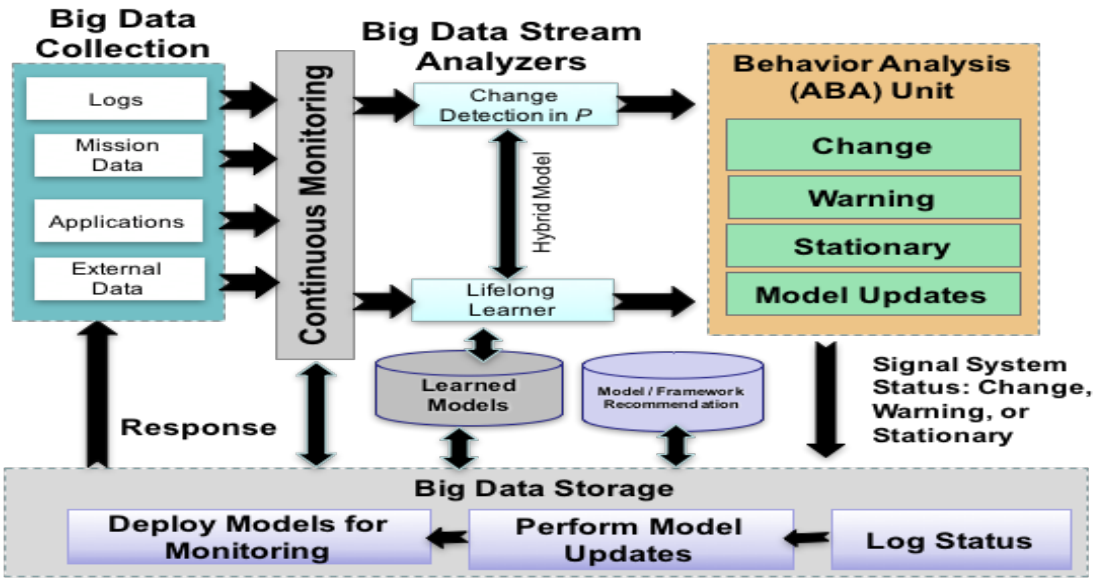


Figure 1. Intelligent data analytics environment.

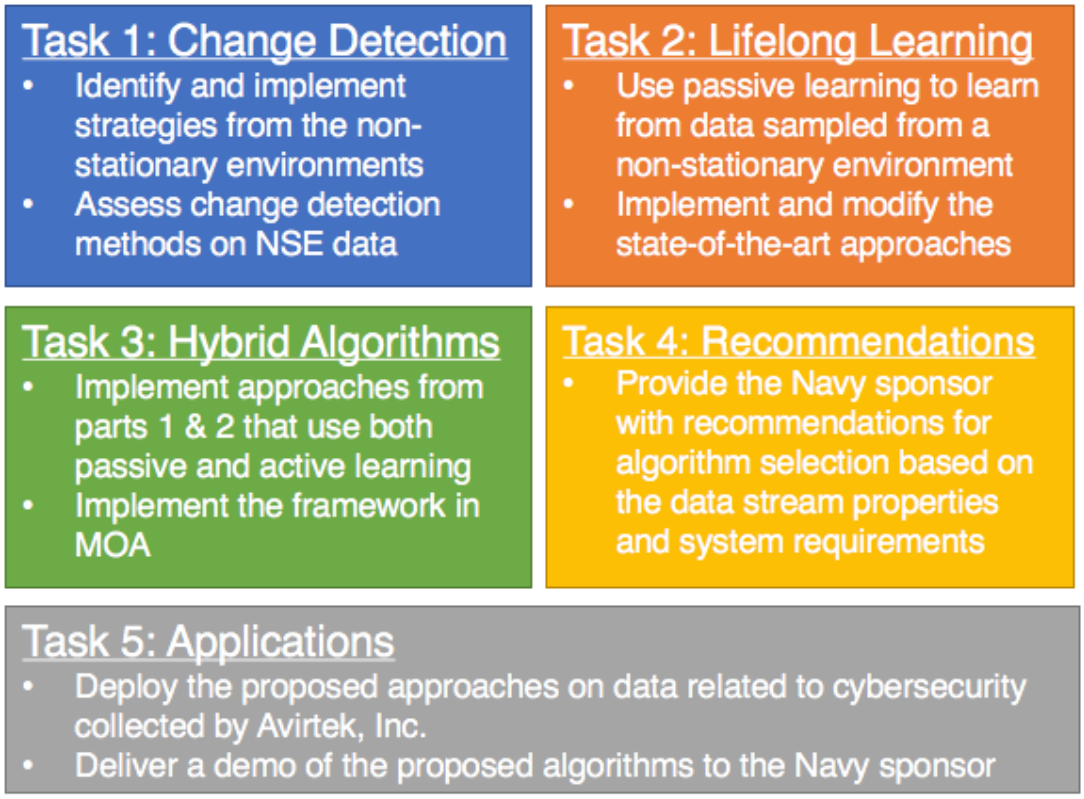


Figure 2. Change detection tasks and applications.

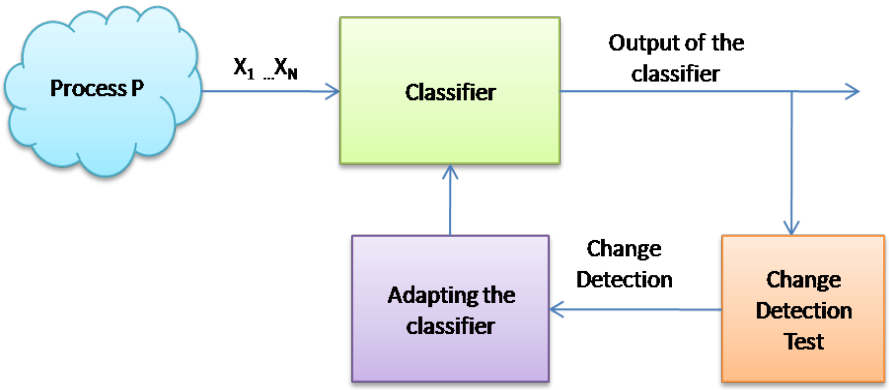


Figure 3. Real-time adaptive machine learning algorithm.

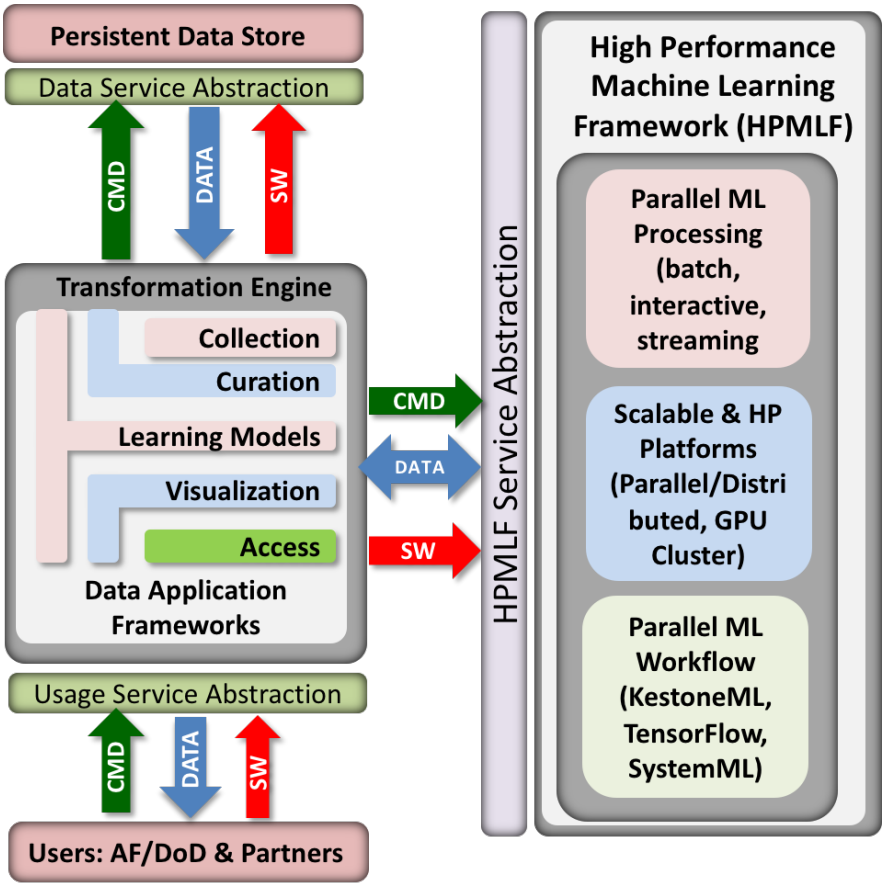


Figure 4. High performance machine learning framework (HPMLF).

Challenges for Smart Cyberinfrastructure in Making Society 5.0 a Reality

Yoshio Tanaka

National Institute of Advanced Industrial Science and Technology, Japan

1. Background

Japan's vision for the future called "Society 5.0" is a super-smart society where technologies such as big data, IoT, AI/ML, and robots fuse into every industry and across all social segments. The concept of Society 5.0 is similar to Cyber Physical Systems (CPS) and there are many challenges to develop smart cyberinfrastructure for making Society 5.0 a reality. One of the biggest challenges is to design the entire edge-to-cloud cyberinfrastructure on which open innovation platforms could be built. These platforms would create an ecosystem of data collection, storage, sharing, and analysis for Society 5.0. In collaboration with universities, research laboratories, and private sectors in Japan and around the world, AIST is meeting this challenge using ABCI (AI Bridging Cloud Infrastructure) as a key resource. ABCI is the world's first large-scale open AI computing infrastructure constructed and operated by AIST. The development of ABCI has the following two phases, (1) develop ABCI as a powerful shared computing infrastructure for AI and (2) develop ABCI as an open innovation platform. Phase

1 is underway since ABCI started its operation on August 2018, it has seen a steady increase in users both from industries and academia. ABCI's software stack and tools provide easy-to-use interfaces for users and achieves highly efficient resource utilization for various types of AI jobs. To achieve Phase 2 of ABCI development, there are many challenges that must be solved through by collaborative research. As a step toward solving these research issues the Japanese government is supporting a data platform project named "mdx" which is being jointly designed by a broad consortium of national universities and institutions.

2. mdx: Data Platform Project

The purpose of mdx is to leverage data utilization throughout Japan making full use of a high-performance research and education network called "SINET". mdx aims to provide a rapid PoC environment for R&D data utilization activities including industry-academia collaboration projects. The infrastructure of the data platform is based on a cloud (IaaS) concept that is distributed over a wide geographical area. Network links connecting data and IoT devices can be provisioned with compute and storage resources securely. The platform provides virtual infrastructure "slices" to users in such a way that the users can use the "slice" as a dedicated secure infrastructure for their purpose. The design of the entire edge-to-cloud cyberinfrastructure is illustrated in Fig. 1. In order to build this data platform, we have to solve

at least the following issues.

(1) Data management

Managing data that enables cross-domain utilization is mandatory and challenging. Data management includes catalogue services and data access mechanisms with appropriate APIs and protocols with many aspects of scalability and high-performance.

(2) Security

Security and privacy which enable the hosting of sensitive data such as medical and personal data is necessary.

(3) Support of real time processing

The data platform will be connected to a huge number of IoT devices directly and must be capable of receiving all data and processing immediately without data loss or leakage.

(4) Job scheduling

Batch scheduling cannot be used for nodes which receive/send real time data, but it is efficient for high machine utilization. The data platform needs to combine long term assigned nodes/VMs with batch scheduled nodes to support jobs.

(5) Management software

IoT devices, wide area network, internal network, computing nodes, storage should be managed in a unified way. Existing cloud management software such as VMware,

OpenStack, Kubernetes, etc. cannot satisfy all the requirements and we have to develop a novel management software.

By solving the above issues, mdx is expected to start its operation in January 2021.

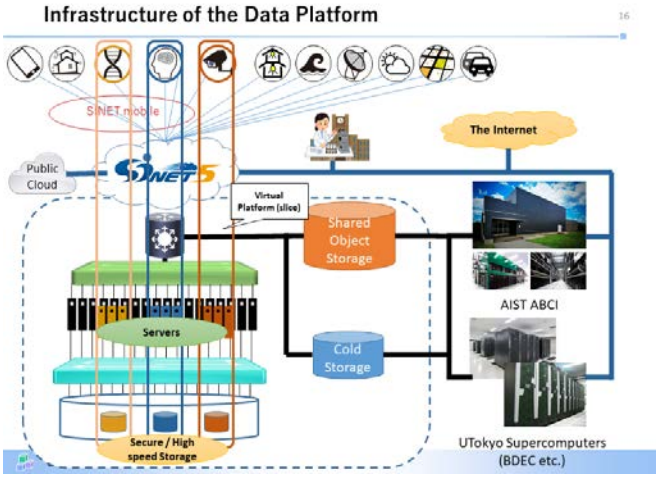


Fig. 1. Infrastructure of the data platform (by courtesy of Prof. T. Kudoh @ U. Tokyo)

Acknowledgement

I would acknowledge to the participants of mdx project.

CyberInfrastructure for Monitoring and Inferencing at the Edge for Sustainable Production

Fatima Anwar, UMass Amherst

1 Problem Statement

Climate change, water shortage, and reduced agricultural land are making the problem of meeting world's increased food demand significantly difficult. Data-driven agriculture practices have benefited farmers in producing crops with reduced water intake [5] [6]. However, agriculture researchers have reported unexpected damage to high value crop such as apples, grapes, and peaches. Major reported cause is variations in environment and pest behaviors due to climate change. Current weather models are coarse and do not capture micro-climates that exist within farms. Agriculture researchers are at the forefront of researching the future of our food production. Researchers are interested in automating agricultural practices that are costly and prone to human error such as fruit pruning and cluster thinning. Also equipping farms to be environment-aware for targeted and reduced chemical spraying. Constant and high-resolution trees monitoring for timely pest management both for disease control and sustainable production are critical for our farm's future. These practices are reliant on the proper use of technology that would enable an in-depth insight into the farm using visual and environmental data at a high *temporal* and *spatial* resolution.

2 Challenges

There has been a lot of work on developing precision agriculture systems, both in industry and academia. However, these systems are application-specific, assume always-on internet connectivity, rely on power sources that negatively impact our environment, does not provide high-resolution spatial and temporal data, does not support real-time applications, and hence do not contribute to our vision of sustainable production. There are numerous challenges in providing an extensible system design to enable precision agriculture for sustainable production. For example, micro-climates affect apple-thinning process and can be captured by high-resolution tree-level sensing capability. Capturing images of apple clusters and correlating them to rich environmental data automatically provides opportune clusters and times to perform apple-thinning without any manual and error-prone process. However such use case demands gathering rich spatial and temporal data that is affected by frequent power outages, intermittent network connectivity, and the availability of low storage at end devices in the farm. Real time use cases also arise i.e. spraying pheromones to attract specific insects and then spraying insecticides at opportune times to reduce chemical exposure. Sensing the right event and taking action at the right time require event-driven and real-time design. To support such use case however, resource-constrained edge sensing devices are incapable of running computationally intensive ML algorithms, while offloading these computations to cloud incur large delays. Finally, various precision agriculture applications in apple-thinning, pest-management, water irrigation, and pollination have different sensing, communication, and actuation requirements. Existing systems cater to only specific applications, for example, See & Spray technology by Blue River focuses only on smart herbicide spraying [1], CropX

platform is designed for smart irrigation and fertilizing [4], and FieldAgent and PrecisionHawk systems focus on generating field maps to monitor crop growth and health [2, 3]. The cost and complexity of combining these systems to provide a holistic view of the farms will deter farmers and academic researchers. To support sustainable production with minimal climate-footprint, it is essential to provide an extensible design to fulfill most precision agriculture applications.

3 Intellectual Merit

The goals of proposed cyberinfrastructure for precision agriculture is increased and sustainable production with reduced cost in terms of error-prone manual labor, and climate foot print. For a transformative landscape in precision agriculture, we foresee various promising research directions. High-resolution spatial and temporal data require an always-on dense deployment of devices that pose challenges in device management, maintenance, and climate footprint of batteries. Open research areas involve designing batteryless devices using energy harvesting mechanisms suitable for agriculture domain. Another area is the design of RFID tags based sensors, and their readers. These designs should ensure system availability even in the face of power and Internet outages caused by bad weather – a fairly common scenario for a farm – resulting in missed communication and so reconstructing missing data is also an open research direction in this field.

Drones are one of the most exciting farm sensors used today [7] but they suffer from poor battery life. Getting aerial imagery for a farm requires multiple drone flights and a long wait time in between when the batteries are being charged. These characteristics are not suitable for our intended applications where we need data of various parts of a single tree every few minutes, i.e. capturing multiple apple clusters on a single tree every 5 minutes. This challenge is addressed by deploying an extensible network and compute architecture with configurable sensing modalities that constantly monitor the entire farm. To support inferencing at the edge to support real-time use cases, an architecture that enables collaborative edge computing by developing novel distributed ML algorithms to satisfy real-time constraints is also a promising research direction in this domain. Finding opportune times and technologies to provide persistent storage over cloud also needs to be explored along with enabling cross-farm analytics.

System Design: For an extensible design, we propose a three-tiered cyberinfrastructure. Lower tier nodes are the most densely deployed and are close to trees and leaves. These devices are batteryless and capable of only sensing and communicating with none to minimal storage. Middle tier nodes are sparsely deployed but are capable of sensing high-resolution data. They receive data from lower tier nodes, store in volatile memory, and process data locally to support sensor fusion, computer vision and ML algorithms at the edge. Higher-tier cloud nodes are capable of persistent storage as well as long-term or cross-farm analytics. This three-tiered architecture open doors for collaborative research across different disciplines and pave path for future research in early disease detection and intervention, pollination patterns, and crop management.

4 Broader Impact

Precision agriculture is a new field that has brought together farmers, diverse academic researchers, and industrial partners to further their common goal of sustainable production. Our proposed architecture does not only address research challenges of immediate adoption but also open doors for future research in studying the impact of climate change on our food supply through early disease control, pest treatment, and artificial pollination. An Interdisciplinary course on precision agriculture will be formulated to train all stakeholders across various disciplines and foster new partnerships.

References

[1] Blue River - See & Spray. http://smartmachines.bluerivertechnology.com/, February 2020.

[2] Drone-based Mapping and Analytics for Agriculture. https://www.precisionhawk.com/agriculture/software, February 2020.

[3] Fieldagent Analytics - Sentera. https://sentera.com/fieldagent-platform/analytics/, February 2020.

[4] The Sensors and Software Behind the Soil Intelligence. https://www.cropx.com/technology/, February 2020.

[5] Mohammed H Almarshadi, Saleh M Ismail, et al. Effects of precision irrigation on productivity and water use efficiency of alfalfa under different irrigation methods in arid climates. *Journal of Applied Sciences Research*, 7(3):299–308, 2011.

[6] Hak-Jin Kim, Kenneth A Sudduth, and John W Hummel. Soil macronutrient sensing for precision agriculture. *Journal of Environmental Monitoring*, 11(10):1810–1824, 2009.

[7] Deepak Vasisht, Zerina Kapetanovic, Jongho Won, Xinxin Jin, Ranveer Chandra, Sudipta Sinha, Ashish Kapoor, Madhusudhan Sudarshan, and Sean Stratman. Farmbeats: An iot platform for data-driven agriculture. In *14th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 17)*, pages 515–529, 2017.

Applying Artificial Intelligence to Broad and Deep Cyber Infrastructure

William Kramer[1], Charles Catlett[2], Alok Choudhary[3], James Brandt[4], Aaron Saxton[1], Saurabh Jha[1], Ann Gentile[4], Ravishankar Iyer[1], Peter Beckman[3], Ewa Deelman[5], Brett Bode[1], Kjellrun Olson[1], Roy Campbell[1], Volodymyr Kindratenko[1], E. A.Huerta[1], Celso Mendes[1], Zbigniew Kalbarczyk[1], Wei-keng Liao[3], Rafael Ferreira Da Silva[5], Ankit Agrawal[3], William Gropp[1]

1 – University of Illinois, 2 – University of Chicago, 3 - Northwestern University, 4 – Sandia National Laboratory, 5 – University of Southern California

Corresponding Author – William Kramer – wtkramer@illinois.edu, +1(217) 979-7577

Abstract: Cyber Infrastructure (CI) is increasingly harnessed and is expanding to support the use of artificial intelligence (AI) techniques and applications to enable smart cities, power grids, computational simulations, real-time processing in big-data experiments (LIGO), supercomputing, and other important innovations. Computation is essential to developing and using the analytics while CI networks and data facilities are central to moving and managing the immense volumes of data that enable progress and insight. Just as AI offers opportunities to revolutionize all aspects of our lives, the potential for improving the design, evolution, optimization, performance, operation, security and sustainability of the CI is equally promising. A convergent research and implementation approach is required to enable the application of *AI for the CI*, comprising experts across the spectrum of cyberinfrastructure research, design, and deployment; computer and computational sciences; and systems design and engineering.

Introduction: All domain science and commerce depend heavily on a robust Cyber Infrastructure (CI) comprising hardware and software systems across ranges of scale (from embedded devices to high performance computing (HPC)), functionalities (from sensing to computational modeling to data analytics to visualization, from data management to communications), and architectures (from centralized to distributed, from tightly coordinated to autonomous, with a wide diversity of underlying technology). All aspects of NSF's, and indeed the nation's, endeavors depend on the creation of scientific and engineering capabilities through the orchestration of CI layers and components that is only increasing as intelligence and autonomy are integrated into all layers and components. Harnessing the power and capacity of current and future CI technologies presents unique and time critical challenges that must be identified, understood, and addressed to maintain and increase the nation's leadership in scientific, industrial, and societal innovation and discovery.

CI is increasingly harnessed to support artificial intelligence[1] (AI) techniques and applications to enable innovation for smart cities, power grids, real-time processing in big-data experiments (such as LIGO), supercomputing simulations, and others. Computation is essential to developing and using the requisite analytics, while CI networks and data facilities are central to acquiring, moving and managing the immense volumes of data required. Just as AI offers promises to revolutionize all aspects of our lives, the potential for improving the design, evolution, optimization, performance, operation, security and sustainability of CI is equally promising. Following the terminology used in the NSF Big Data and Extreme Computing Workshop series we refer to AI applied to challenges of improving Cyber Infrastructure as AI for CI [1].

A convergent research approach is required to explore the application of AI for CI, comprising experts from across the spectrum of cyberinfrastructure research, design, and deployment; computer and computational sciences; and systems design and engineering. To this end, we propose to create a National AI for CI research and development effort to drive transformative exploration of CI using AI methods including areas ranging from CI performance and optimization to the concept of enabling scientists to specify high-level science objectives from which workflows might be constructed and resources (hardware, software, data) might be assigned and made available.

There are clear needs for AI innovations and implementations focused on the CI itself to build a new multidisciplinary community of AI and CI researchers. It will include people designing, implementing and supporting CI across the entire CI spectrum (from the largest to the smallest systems, from the tightest to loosest integration); and core AI/ML experts who can explore the use of AI to improve the effectiveness and efficiency of CI design and implementations.

[1] while there are differences in implementations and use of Machine Learning, Deep Learning and Artificial Intelligence, there are related for the sake of this paper we will refer to all methods as AI

Conceptual Framework: Harnessing AI to Reinvent Cyber Infrastructure: Machine Learning (ML), Deep Learning (DL), and AI have recently become important and practical alternative methods for data-driven discovery in a number of science and engineering domains including physics, agriculture, natural language processing, facial and object recognition, imaging, and automation. Recent successes have been enabled a combination of massive increases in raw processing power, the availability of exponentially increasing data for analysis and learning and successes in algorithms for applying that processing power to solving AI problems. All of these areas now rely on powerful and stable CI to apply AI methods to accomplish their domain science goals and innovation.

These examples of using CI to empower AI represent tremendous research opportunities [5]. CI is complex, vast, expensive, and can be difficult to use effectively. AI for CI endeavors will help reduce errors, decrease the time needed to rectify problems, faults and inefficiencies, and significantly decrease barriers to improving AI and other application performance.

Disease vectors are increasing at an alarming rate with the increase in travel and export/import on a global scale as evidenced by the two recent outbreaks in coronavirus and aggressive flu epidemics. Creating timely cures and modeling contagion spread requires sophisticated CI that is currently lacking (e.g., distributed sensors coupled with computation, global connections of medical providers, etc.) All aspects of enabling this acceleration hinge on our ability to efficiently build and harness the CI that is integral to all aspects of modern technology. This is but one example of current and urgent problems; improving the CI addresses many other science, research, and engineering innovations needs that will enable acceleration of productivity and decreased time to insight/solution across all disciplines. Enhancing prediction of severe weather, reducing time to discovery of new drugs and new materials, understanding our changing earth and these changes impacts on society and security, reducing energy use and increasing efficiency, understanding human behavior, understanding the universe's origin and function, and many other areas all rely on fundamentally and dramatically improving how our CI works. With the rate for hardware enabled improvements decreasing we must increasingly rely on much smarter and more effective software.

To enable the end goal of accelerating discovery, it is critical that we incorporate AI improvements in our CI starting now. The urgency comes from two parallel time-critical issues. First, AI for CI can enable important improvements to the existing and soon to be deployed CI over the next 5 years but only if we begin work immediately. Possibly even more important, new, long lasting CI architectures and systems are being designed today and it is imperative to their future performance that the new architectures incorporate AI in their CI in a comprehensive and efficient manner from the ground up.

The Human-in-the-Loop (HIL) methods prevalent in many aspects of CI operations today are too limited when dealing with complexities for even small CI implementations and may become impossible when complexity grows to national scale CI. The cost of HIL in terms of inefficient use, delayed or incomplete discovery, or simply unavailability of CI components is now too high given CI is expensive and each implementation has a limited time-span window to make an impact. The current "top-down, human-in-the-loop" approaches for its design, implementation, effective use and improvement are very limiting and prone to error. Studies [7] [8] show that human actions contribute significantly to downtime, errors, and rework in CI implementations. HIL is typically limited to being reactive (or retrospective), using a postmortem analysis of events, anomalies or performance. Data-driven AI based techniques appear the best options for making predictive approaches possible.

Ultimately, AI has opportunity to go beyond simply assisting or augmenting present CI capabilities. It will be used to explore new CI capabilities such as developing high-level scientific problem specification constructs that can discover and evaluate relevant resources (data sources, instruments, computation) and services (algorithms, models), testing promising workflows from such components, and providing end users with prototype workflows to explore a specified problem.

How might AI approaches be used to address CI challenges? First, AI can reduce the complexity of designing and operating CI. A single and/or small system may consist of millions of interacting components, whereas a larger, national-class system may contain billions of components. Similarly, the amount of code used to implement the systems ranges from millions to billions of lines. By definition, CI is highly interconnected and systems are co-dependent. This complexity makes systems that afford security and sustainability increasingly difficult to correctly design, build, and operate. Discovering services and optimizing their interaction in workflows in CI will require new, more scalable approaches than in the past.

Second, AI can reduce the delay in responding to conditions of interest. Almost all CI systems have significant limitations for data movement. To improve efficiency, workflows and overall system performance, mismatches responses to changes in conditions and delays in mitigation generation must be identified and alleviated. With the increased power of edge devices and the heterogeneity of network capabilities, emerging CI systems require complex orchestration of computation and data movement to reduce dependence on communications (e.g., reducing data volume) as well as to enable CI to act in real time despite having components that are tens of milliseconds or further apart. Concurrently, the use of embedded, wearable, or other personal devices expands the concept of CI.

Third, AI can reduce "Time-to-discovery" and "Time-to-actionable insights" in CI operations. Data sizes and complexities limit human's ability to discover anomalies, inefficiencies, bottlenecks, faults, security vulnerabilities, and cyber-attacks on time-scales appropriate for taking corrective actions. AI methodologies can efficiently extract such insights from large, complex data. However, as CI workflows and services become more complex it will be essential to improve our ability to create systems that behave in a way that can be explained to ensure confidence in taking action upon AI-based results.

Specifically, CI that is used across applications that require resources ranging from the farthest edge devices to the largest of facilities, from dedicated ultra-scale instruments to the broad range of distributed devices, from stationary to mobile and autonomous devices and systems. Brief review of a few exemplar use cases help explain the challenges and objectives of AI for CI, adding new use cases as more insight is gained. These exemplars include (i) mid- and large-scale data and computational facilities, (ii) CI for large scale observing instruments, (iii) CI for automated and robotic devices, and (iv) CI for widely distributed instruments whether it is in cities or national scale. Social science considerations for CI are shared among these cases. Each exemplars use has unique challenges but it is also possible to identify commonality that can be applied to all or are easily adapted for different CI implementations. We call these exemplars "Families of CI". These similarities will be conducive to developing common algorithmic approaches, interfaces, and protocols which will likely also extend to many other areas including those necessary in complex industrial scenarios. These families represent opportunities for AI for CI but are not exhaustive, and indeed are expected to be expanded.

Research Themes for Exemplar Families of Scientific CI: *Computational and Data Systems*: CI for computational and data analysis systems and facilities come in all sizes from a modest handful of compute nodes with a Network File System (NFS) to support common storage to complex High Performance Compute (HPC) systems consisting of tens of thousands of compute nodes and tightly integrated subsystems for supporting large scale data buffering and file I/O, to global scale cloud computing facilities. Across the spectrum of size, configuration and workload, C&D CI systems share the challenge of being used inefficiently. Application developers struggle determining the best algorithms to optimally utilize the increasing diversity of architectural features. These challenges are only getting worse with the ever-increasing complexity of all compute, network and data technologies. Across all such facilities this not only translates into waste in terms of energy, idle resources, and human capital, but also restricts the rate and quality of insight and output and extends the time to solution of some of humanity's most pressing problems.

In order to optimize the utilization of our compute and data related resources and minimize the time to solution of the problems being run on these resources, the technology industry and consumers have been adding CI in the form of instrumentation, communication, and analysis infrastructure for collecting and analyzing detailed information about all aspects of operation. To date this additional CI increases the burden on support staff and users of the technology to understand the complex interactions represented by the information and has been used largely for diagnosing problems reported and performance bottlenecks. Due to the high volume (tens of TB/day per system) and high dimensionality (thousands of different instrumentation points) of data, humans are only able to use a small subset of the data for analysis and are often unable to identify root causes of problems. Additionally, delays between problem occurrence and diagnosis for all but the most trivial problems limit the utility of current CI systems with respect to automated feedback and optimization of run time resources and configuration [10].

Research Themes for Exemplar Families of Scientific CI: *CI for Large Instruments:* The CI demands for next generation instruments, such as the Large Synoptic Survey Telescope (LSST) [3], are expected to exceed existing CI capabilities [11]. Upgrades to existing facilities, e.g., the advanced Laser Interferometer Gravitational-wave Observatory (LIGO) [12], and the next generation Large Hadron Collider [13] already demand the use of CI that are beyond in-house, tailored solutions for core-data analyses, requiring the use of the Open Science Grid, and other HPC platforms such as XSEDE and Blue Waters [14] [15]. There is

an urgent need to rethink the existing CI paradigm to cope with the volume and speed of data production and analysis of large-scale facilities.

The convergence of AI with cloud, mid- and top-tier HPC centers has rapidly emerged as an alternative to handle the disruptive effects of the big-data revolution in strategic NSF- and DOE-funded investments. While the design and training of AI algorithms requires novel cyber-tools, such as distributed training schemes in HPC platforms, once these algorithms are fully trained, they enable real-time inference studies using minimal resources. Convergence between existing CI and AI-HPC inspired CI solutions, with expanded AI for CI technology, could boost and enrich the scientific capabilities of current and next-generation big-data experiments [16]. This approach will leverage the expertise of system architects, resource managers, and software developers to use AI to automate diagnostics tools that optimize the performance and throughput of CI resources. AI for CI will also carry out data-driven analyses, informed by the high-dimensionality of the in-situ system and application telemetry data, to form logical hypotheses about the data features essential to understanding relationships among architectures, behavioral characteristics of applications, and performance, even in the absence of labelled associated application runs and conditions, alleviating the need for relying on limited benchmarks and regression testing that is common practice today.

Research Themes for Exemplar Families of Scientific CI: *CI for Automation and Robotics*: AI based cyber-physical autonomous agents that directly assist or interact with humans are becoming ubiquitous [17] [18] [19] [20]. However, field-deployment of trustworthy, safe, and performant autonomous agents has the following two CI interrelated challenges. First, the lack of automated processes for design and development of an end-to-end field-deployable autonomous agent CI [21] [22]. Addressing the system challenges in design and development is critically important as evident from the delays in deployment of self-driving cars despite the showcase of a working concept of a functional AV in DARPA 2007 Urban Challenge [23] 13 years ago. Towards this goal, new and innovative automated processes must be created for operating systems and computational kernels that meet energy, latency and deadline requirements, and methods/protocols for testing, failure mitigation, recovery, and safety. Moreover, these problems are increasingly more challenging in the context of autonomous swarms that are geared towards execution of shared tasks (e.g., search and rescue missions) as they require methods for remote management (from a base station) and coordination of geo-distributed agents (edge devices) in uncertain human-centric environments. The second challenge is the lack of data, code, and knowledge sharing across the field of automation and robotics [24] [25]. There is a significant commonality across autonomous agents in the field of automation and robotics, such as self-driving vehicles, manufacturing/industrial/surgical robots, and delivery bots. This commonality spans CI requirements and specifications (e.g., remote management and coordination), meeting real-time deadlines for executing actions, and maintaining safety. Although agents operate at different levels of autonomy and perform significantly different tasks (e.g., human transportation vs. package delivery), the significant commonality among them can be leveraged by investment in standards to create reusable design, development, and deployment patterns to maximize safety, efficiency, and benefit to society.

In order to test ML and AI algorithms for perception, decision, and control of autonomous agents, industry has developed several open-source comprehensive testbeds [26] and [27] that utilize real-time physics-based simulators to support software-in-the-loop, hardware-in-the-loop, and human-in-the-loop methods. This allows developers to iterate on architecture, design and implementation of the autonomous CI system and generate large amounts of data on execution time of computational tasks, their safety, and resilience.

Unique AI opportunities for automation/robotics: AI approaches are promising as the current methods for design and development of safe CI for robotics and automation are fragile as it depends on painstakingly tuned heuristics and partial solutions with insufficient safety assurance. AI-techniques can leverage simulation and field datasets as well as domain knowledge to enable automatic design space exploration and development of CI components that are performant, safe and resilient.

Research Themes for Exemplar Families of Scientific CI: *Edge Systems and Distributed Intelligent Observatories*: NSF has supported the creation of distributed observatories ranging in scale from urban (e.g., Array of Things (AoT) in Chicago [28]) to regional (e.g., WIFIRE in Southern California [29]) to continental (e.g, National Ecological Observatory Network (NEON) [30]). At each of these scales there is need for measurements outside of the capabilities of traditional sensors, aimed at sensing factors ranging from vehicle or pedestrian flows in cities or the migratory patterns of wildlife. Desired capabilities include the ability to modify sampling rates to suit current conditions (e.g., low sampling rates when nothing out of

the ordinary is going on and higher sampling rates surrounding events or conditions of interest that resolve desired detail) and the ability to aggregate and anonymize data to levels appropriate to types of information being collected (e.g., anonymous pedestrian and traffic flows vs. identification in the case of collisions and altercations).

Industry has begun a transition to monitoring and control using Internet-of-Things (IoT) technologies and is rapidly adopting edge computation to embed AI algorithms. Whether in factories, electrical and gas distribution networks, or urban measurement instruments, edge computation supports the creation of "software-defined sensors" that, like software-defined networks, are remotely programmed to allow flexibility and evolution of their measurement capabilities.

Unique AI opportunities for edge systems include: AI-at-the-edge can provide support for multiple types of CI needs, for instance enabling instruments to enhance communication conserving low-frequency samples with anomaly detection to trigger changes in sampling rates appropriate to resolve events or conditions of interest. For example, cameras deployed in cities need not stream and save (with significant privacy implications) video to central servers if their embedded intelligence can both extract routine measurements (e.g., pedestrian flows) and adapt to events of interest (such as a vehicle collision). As new CI capabilities move from prototypes to production scale, CI services in observatories such as NEON [31], AI will be increasingly critical to the operation and optimization of distributed components with high degrees of heterogeneity, diverse services, and in some cases intermittent connectivity.

Social Science Influence on CI: The social, economic, and behavioral science communities have significant experience working with sensitive data, maybe more than do the other NSF-supported sciences and have grappled with the ethics of the use of "big data" [32]. The advent of smartphones, wearables, in-home devices, and location-based services has brought CI squarely into social science research. The pace and scale of the use of such CI in the social sciences are such that the privacy, security, and ethics of the CI, and moreover of the use of AI within these devices, require a central rather than supporting role for social science in the use of AI for CI.

Equally important is the impact of AI for CI from the point of view of stakeholder trust--whether the correctness of results, the security of systems embedded in the public way and in homes, or the ethical application of AI in the use of associated data. Without trust, there will be no adoption. Initiatives must address increasing autonomy in order to reduce latency for making high-quality decisions in high-consequence and/or complex scenarios with the goal of increased performance, efficiency, security and safety. CI is a place where domain scientists make high consequence decisions and getting those decision-makers and their stakeholders to trust AI is going to be difficult unless its explainable and transparent in some way. In particular, there is a need to explore the partnership and trust between the AI-decision making tools and the humans for high risk tasks.

AI for CI research areas include low- through high-risk scenarios. Low-risk scenarios without significant incoming bias can provide a good testbed for testing acceptance and assessing confidence in results. For example, HPC scheduling decisions to improve performance can be low-risk as long as the affected applications are not high-consequence. Assessing criteria for tradeoffs in explainability-transparency-acceptance-latency change higher risk scenarios (e.g., more urgent computations, autonomous vehicles) will be critical to acceptance of the new methods.

Common Challenge Themes: The research challenges identified across the exemplar CI families all share some common themes: 1) CI systems are so complex that only domain experts can make reasonable and repeatable optimization decisions and the complexity is growing exponentially, 2) decision making should be performed with minimal lag time for optimum performance but the consequences of those decisions, in terms of human welfare and/or monetary cost, can be high, 3) processing telemetry data is a "big data" problem in itself with requirements for low latency time to solution in the presence of perhaps incomplete data, 4) individual data values are bounded but the high dimensionality and variability in aggregate makes full ensemble processing, within effective latency bounds, currently impossible, 5) social aspects of trusting the judgment of a machine must be overcome for significant deployment, 6) workload and resource scheduling, swarm coordination/management, resiliency, energy efficiency, ease of use by developers and end user, and 6) lack of labeled datasets as well as the high-cost of generating labeled datasets for training.

Across CI families, AI approaches will enable domain experts to automate the design space exploration of performant, safe and reliable CI, and unlock new use across the domains of science, health, agriculture, manufacturing, energy, and transportation by combining field datasets, simulation techniques and domain

knowledge. We believe that these similarities will be conducive to the development of common algorithmic approaches (including explainability), interfaces, and protocols which will likely also extend to areas such as weather forecasting using drones/aircraft and more.

Conclusion: A National AI-for-CI effort that organizes a community of both CI and AI experts and stakeholders to focus on identifying the challenges and opportunities for AI to enhance and eventually re-invent the CI is a critical national need. Equally important is AI for CI efforts including leaders from some of the largest and most complex of NSF's CI projects and operating observatories and facilities to bring a diverse set of stakeholders and experts together from both a disciplinary standpoint and a human standpoint. The ubiquitous and foundational changes to be empowered by applying AI to CI require new perspectives from new communities that challenge traditional thinking embedded in nearly four decades of NSF-supported CI. Simply put, only such a bold approach can yield insights and new concepts that are required to fully exploit AI in achieving the promise of increasingly broad, powerful, and intelligent CI in the coming decade

REFERENCES CITED

[1] BDEC2, "Big Data and Extreme-Scale Computing 2," 2019. [Online]. Available: https://www.exascale.org/bdec/meeting/sandiego. [Accessed 27 January 2020].

[2] G. H. Bauer, B. Bode, J. Enos, W. T. Kramer, S. Lathrop, C. L. Mendes and R. R. Sisneros, "Best Practices and Lessons from Deploying and Operating a Sustained-Petascale System: The Blue Waters Experience," in *Proceedings of Supercomputing'18*, Dallas, TX, 2018.

[3] LSST, "The Large Synoptic Survey Telescope," [Online]. Available: https://www.lsst.org/lsst. [Accessed 27 January 2020].

[4] MBDH, "Midwest Big Data Hub," [Online]. Available: http://midwestbigdatahub.org/. [Accessed 27 January 2020].

[5] DOE, "AI for Science Town Hall Meetings," 2019. [Online]. Available: https://www.ornl.gov/content/town-hall-artificial-intelligence. [Accessed 27 January 2020].

[6] G. Fox, "Perspectives on High-Performance Computing in a Big Data World," in *Proceedings of the 28th International Symposium on High-Performance Parallel and Distributed Computing*, Phoenix, AZ, 2019.

[7] A. Brown and D. A. Patterson, "To Err is Human.," in *Proceedings of the First Workshop on Evaluating and Architecting System DependabilitY (EASY '01),*, Gothenburg, Sweden, July 2001..

[8] A. Brown, L. C. Chung and D. A. Patterson, "Including the Human Factor in Dependability Benchmarks," in *Proceedings of the DSN Workshop on Dependability Benchmarking*, Washington, DC, 2002.

[9] "Digital Twin," 2 January 2020. [Online]. Available: https://en.wikipedia.org/wiki/Digital_twin.

[10] S. Jha, J. Brandt, A. Gentile, Z. Kalbarczyk, G. Bauer, J. Enos, M. Showerman, L. Kaplan, B. Bode, A. Greiner, A. Bonnie, M. Mason, R. K. Iyer and W. Kramer, "Holistic Measurement-Driven System Assessment," in *Proceedings of the IEEE International Conference on Cluster Computing (CLUSTER)*, Honolulu, HI, 2017.

[11] NAS, "Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science and Engineering in 2017-2020," National Academies of Sciences, Engineering, and Medicine, Washington, DC, 2016.

[12] LIGO, "Laser Interferometer Gravitational-wave Observatory," [Online]. Available: https://www.ligo.caltech.edu/page/what-is-ligo. [Accessed 27 January 2020].

[13] D. Castelvecchi, "Next-generation LHC: CERN lays out plans for €21-billion supercollider," *Nature,* Vols. 565, 410 (2019), 15 January 2019.

[14] E. A. Huerta, R. Haas, S. Jha, M. Neubauer and D. S. Katz, "Supporting High-Performance and High-Throughput Computing for Experimental Science," *Computing and Software for Big Science,* vol. 3, no. 5, 2019.

[15] E. A. Huerta, R. Haas, E. Fajardo, D. A. Katz, S. Anderson, P. Couvares, J. Willis, T. Bouvet, J. Enos, W. T. Kramer, H. W. Leong and D. Wheeler, "BOSS-LDG: A novel computational framework that brings together Blue Waters, Open Science Grid, Shifter and the LIGO Data Grid to accelerate gravitational wave discovery," in *Proceedings of the 13th IEEE International Conference on e-Science*, Auckland, 2017.

[16] E. Papka, I. Monga, J. J. Hack and S. Jones, "Facilities Integration and AI Ecosystem," [Online]. Available:

https://www.csm.ornl.gov/documents/downloads/ai/Facilities%20Integration%20and%20AI%20Ecosystem/FacilitiesIntegrationORNLOutbriefV2.pptx. [Accessed 27 January 2020].

[17] F. Codevilla, M. Müller, A. López, V. Koltun and A. Dosovitskiy, "End-to-End Driving Via Conditional Imitation Learning," in *Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA)*, Brisbane, 2018.

[18] A. Hussein, M. M. Gaber, E. Elyan and C. Jayne, "Imitation Learning: A Survey of Learning Methods," *ACM Computing Surveys (CSUR),* vol. 50, no. 2, pp. 1-35, 2017.

[19] ResearchAndMarkets, "Global Agriculture Drones and Robots Markets, 2018-2028 - Advancements in Artificial Intelligence Technology and Machine Learning Solutions," 2019. [Online]. Available: https://www.globenewswire.com/news-release/2019/05/10/1821686/0/en/Global-Agriculture-Drones-and-Robots-Markets-2018-2028-Advancements-in-Artificial-Intelligence-Technology-and-Machine-Learning-Solutions.html. [Accessed 27 January 2020].

[20] J. Edward, "Building a Smart Factory with AI and Robotics," 2018. [Online]. Available: https://www.roboticsbusinessreview.com/wp-content/uploads/2018/02/RBR_BuildingAI_WP3.pdf. [Accessed 27 January 2020].

[21] L. E. Lwakatare, A. Raj, J. Bosch, H. H. Olsson and I. Crnkovic, "A Taxonomy of Software Engineering Challenges for Machine Learning Systems: An Empirical Investigation," in *Proceedings of the International Conference on Agile Software Development*, Montreal, 2019.

[22] B. Hrvoje, M. Vukovic and C. Željka, "Requirements Engineering Challenges in Building AI-Based Complex Systems," in *Proceedings of the 27th IEEE International Requirements Engineering Conference Workshops (RE'19)*, Jeju Island, South Korea, 2019.

[23] DARPA, "DARPA 2007 Urban Challenge," 2007. [Online]. Available: https://www.darpa.mil/about-us/timeline/darpa-urban-challenge. [Accessed 27 January 2020].

[24] F. L. DaSilva and A. H. Costa, "A survey on transfer learning for multiagent reinforcement learning systems," *Journal of Artificial Intelligence Research,* vol. 64, pp. 645-703, 2019.

[25] C. Longbing, G. Weiss and S. Y. Philip, "A brief introduction to agent mining," *Autonomous Agents and Multi-Agent Systems,* vol. 25, no. 3, pp. 419-424, 2012.

[26] AirSim, "Microsoft/AirSim," Microsoft, [Online]. Available: https://github.com/microsoft/AirSim. [Accessed 27 January 2020].

[27] CARLA, "Open-source simulator for autonomous driving research," [Online]. Available: http://carla.org. [Accessed 27 January 2020].

[28] C. E. Catlett, P. H. Beckman, R. Sankarn and K. K. Galvin, "Array of things: a scientific research instrument in the public way: platform design and early lessons learned," in *Proceedings of the 2nd International Workshop on Science of Smart City Operations and Platforms Engineering (SCOPE'17)*, 2017.

[29] I. Altintas, L. Smarr, H. W. Braun and R. D. Callafon, " Hazards SEES Type 2: WIFIRE: A Scalable Data-Driven Monitoring, Dynamic Prediction and Resilience Cyberinfrastructure for Wildfires," NSF Award 1331615, 2013. [Online]. Available: https://www.nsf.gov/awardsearch/showAward?AWD_ID=1331615. [Accessed 27 January 2020].

[30] D. Schimel, W. Hargrove, F. Hoffman and J. MacMahon, "NEON: A hierarchically designed national ecological network," *Frontiers in Ecology and the Environment,* vol. 5, no. 2, pp. 59-59, 2007.

[31] NEON, "The National Ecological Observatory Network," [Online]. Available: https://www.neonscience.org. [Accessed 27 January 2020].

[32] P. Elias, P. H. Fossheim, P. Muhlberger, C. Catlett, K. A. Cagney, R. George, G. v. Halderen, T. Ausin, K. Agni, S. Maurizio, S. Hagit, P. Buleon, S. Bender, M. Martin, I. Satoh, S. Osaumu, M. Flesaoana, C. v. Zyl, M. Woolard, S. McGregor and C. Humphrey, "Research ethics and new forms of data for social and economic research.," *OECD,* no. 34, 2016.

[33] P. Beckman, C. Catlett, I. Altintas, S. Collis and E. Kelly, "Mid-Scale RI-1: SAGE: A Software-Defined Sensor Network," NSF Award 1935984, 2019. [Online]. Available: https://www.nsf.gov/awardsearch/showAward?AWD_ID=1935984&HistoricalAwards=false. [Accessed 27 January 2020].

[34] DES, "The Dark Energy Syrvey," [Online]. Available: https://www.darkenergysurvey.org. [Accessed 27 January 2020].

[35] NCSA, "Center for Artificial Intelligence Innovation," University of Illinois, [Online]. Available: http://www.ncsa.illinois.edu/about/centers/ai/. [Accessed 27 January 2020].

[36] N. Stout, "Supercomputers will start building a 3D map of the world," 5 August 2019.

[37] S. Jha, S. S. Banerjee, T. Tsai, S. K. Hari, M. Sullivan, Z. Kalbarczyk, S. Kecler and R. K. Yyer, "ML-based Fault Injection for Autonomous Vehicles: A Case for Bayesian Fault Injection," in *Proceedings of the 49th IEEE/IFIP International Conference on Dependable Systems and Networks*, Portland, 2019.

[38] R. K. Iyer, A. P. Athreya, L. Wang and R. M. Weinshilboum, "Artificial Intelligence and Pharmacogenomics: A Timely Synergy for Individualizing Medicine," *Advances in Molecular Pathology,* vol. 2, no. 1, pp. 111-118, 2019.

[39] A. P. Athreya, A. J. Gaglio, J. Cairns, K. R. Kalari, R. M. Weinshilboum, L. Wang, Z. T. Kalbarczyk and R. K. Iyer, "Machine Learning Helps Identify New Drug Mechanisms in Triple-Negative Breast Cancer," *IEEE Transactions on NanoBioscience,* vol. 17, no. 3, pp. 111-118, 2018.

[40] Y. Varatharajah, B. Berry, J. Cimbalnik, V. Kremen, J. V. Gompel, M. Stead, B. Brinkmann, R. Iyer and G. Worrell, "Integrating artificial intelligence with real-time intracranial EEG monitoring to automate interictal identification of seizure onset zones in focal epilepsy," *Journal of Neural Engineering,* vol. 15, no. 4, 2018.

[41] S. S. Banerjee, R. Jha and R. K. Iyer, "Inductive Bias-driven Reinforcement Learning For Efficient Schedules in Heterogeneous Clusters," 2019. [Online]. Available: https://arxiv.org/abs/1909.02119. [Accessed 27 January 2020].

[42] S. Jha, S. Cui, T. Xu, J. Enos, M. Showerman, M. Dalton, Z. T. Kalbarczyk, W. T. Kramer and R. K. Iyer, "Live Forensics for Distributed Storage Systems," 2019. [Online]. Available: https://arxiv.org/abs/1907.10203. [Accessed 27 January 2020].

[43] V. Satone, R. Kaur, H. Leonard, H. Iwaki, L. Sargent, S. Scholz, M. A. Nalls, A. B. Singleton, F. F. Faghri and R. H. Campbell, "Predicting Alzheimer's disease progression trajectory and clinical subtypes using machine learning," 2019. [Online]. Available: https://doi.org/10.1101/792432. [Accessed 27 January 2020].

[44] S. Hashemi, S. Yothi and R. Campbell, "TicTac: Accelerating Distributed Deep Learning with Communication Scheduling," in *Proceedings of the SysML Conference*, Stanford, CA, 2019.

[45] A. Agelastos, B. Allan, J. Brandt, P. Cassella, J. Enos, J. Fullop, A. Gentile, S. Monk, N. Naksinehaboon, J. Ogden, M. Raian, M. Showerman, J. Stevenson, N. Taerat and T. Tucker, "Lightweight Distributed Metric Service: A Scalable Infrastructure for Continuous Monitoring of Large Scale Computing Systems and Applications," in *Proceedings of Supercomputing'14*, New Orleans, 2014.

[46] O. Tuncer, E. Ater, Y. Zhang, A. Turk, J. Brandt, V. J. Leung, M. Egele and A. K. Coskun, "Diagnosing Performance Variations in HPC Applications Using Machine Learning," in *Lecture Notes in Computer Science - High Performance Computing ISC*, Springer, Cham, 2017, pp. 355-373.

[47] CUCIS, "Center for Ultra-scale Computing and Information Security," Northwestern University, [Online]. Available: http://cucis.ece.northwestern.edu/. [Accessed 27 January 2020].

[48] "JELSC Overview," 29 January 2020. [Online]. Available: https://jlesc.github.io/.

[49] GE, "Digital Twin Framework," General Electric, [Online]. Available: https://www.ge.com/research/project/digital-twin-framework. [Accessed 27 January 2020].

Advanced Cyberinfrastructure for Accelerating Science[1]

A whitepaper submitted to the NSF Workshop on Developing a Roadmap towards the Next Generation of Smart Cyberinfrastructure

Vasant G. Honavar[2]
Artificial Intelligence Research Laboratory
Center for Big Data Analytics and Discovery Informatics
Institute for Computational and Data Sciences
Pennsylvania State University

Tycho Brahe gathered considerable and accurate data on the movement of the planets ("big data" for his time). However, this data did not find real value until Johannes Kepler used it to discover his three laws of planetary motion. Later Isaac Newton used these laws and other data to derive his unified laws of motion and laid the foundations of classical physics. To do so, he had to invent calculus for describing such things as rates of change. Brahe, Kepler, and Newton were all engaged in the practice of science, a systematic process for acquiring knowledge through observation or experimentation and developing theories to describe and explain natural phenomena. The scientific process that they engaged in is summarized in Figure 1.

Typically, scientific inquiry starts with a question within a domain of study, e.g., biology. With the question in hand, one has to assemble the background information and acquire the data necessary to answer the question. Then one proceeds to construct one or more models from data (and background information). Choosing a small set of models from among a much larger set of candidates involves additional

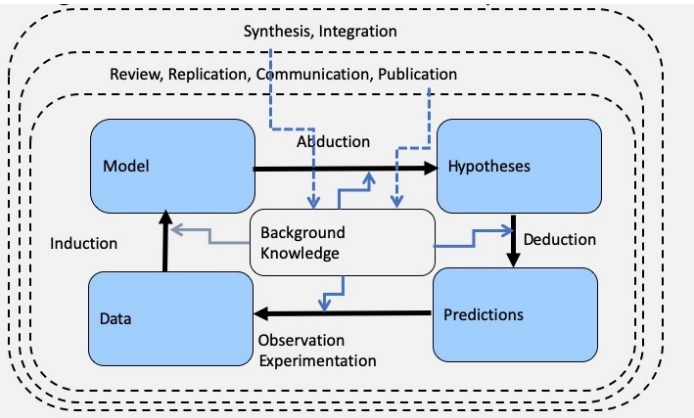


Figure 1: Major Components of the Scientific Process

[1] Some of the material in this white paper has been adapted from: (i) Honavar V., Hill M., & Yelick K. (2016). *Accelerating Science: A Computing Research Agenda*. A white paper prepared for the Computing Community Consortium committee of the Computing Research Association. arXiv:1604.02006 and (ii) Honavar V., Yelick K., Nahrstedt K., Rushmeier H., Rexford J., Hill M., Bradley E., & Mynatt E. (2017). Advanced Cyberinfrastructure for Science, Engineering, and Public Policy. A white paper prepared for the Computing Community Consortium committee of the Computing Research Association. arXiv:1707.00599

[2] Professor and Edward Frymoyer Chair of Information Sciences and Technology; Professor, Computer Science, Bioinformatics and Genomics, Informatics, and Neuroscience Graduate Programs; Professor, Data Sciences Undergraduate Program; Director, Artificial Intelligence Research Laboratory; Director, Center for Big Data Analytics and Discovery Informatics; Associate Director, Institute for Computational and Data Sciences; Co-PI, Northeast Big Data Innovation Hub; Co-PI, Virtual Data Collaboratory; Steering Committee Member, Eastern Regional Network

considerations (simplicity, consistency with what else is known), etc. The models can be used to advance hypotheses that result, ideally, in testable predictions. The observations or experiments designed to test the predictions yield additional data that feed into the larger scientific process. Science is a social endeavor, with multiple individuals and teams, driven by intrinsic as well as extrinsic incentives. Scientific findings go through peer review, communication, and publication, and replication before they are integrated into the larger body of knowledge in the relevant discipline. It is worth noting that there is considerable variability across scientific disciplines, e.g., in cosmology, where there is little possibility of executing designed experiments, one typically has to make do with observational data or the results of 'natural' experiments. Nevertheless, it is clear that the processes of acquiring, organizing, verifying, validating, integrating, analyzing, reasoning with, and communicating information (models, hypotheses, theories, explanations) about natural and built systems lie at the heart of the scientific enterprise. The past centuries have witnessed major scientific breakthroughs as a result of advances in instruments of observation, formalisms for describing the laws of nature, improved tools for calculation, and infusion of concepts, tools, and scientific practices across disciplines.

Today, the experimental instruments are more powerful, the scientific questions more complex, and the mathematical, statistical and computational methods for analyzing data have become more sophisticated. The resulting emergence of "big data" offers unprecedented opportunities for accelerating science. Arguably, "big data" accelerates Brahe's part of the scientific endeavor, and increasingly, Kepler's part, with the increasing use of machine learning for building models from data. **Nevertheless, most other aspects of the scientific process (understanding the current state of knowledge, formulating questions, designing studies, assembling and managing research teams, identifying designing, prioritizing, optimizing and executing experiments, organizing and integrating data, knowledge, and assumptions to draw inferences and interpret and explain results) constitute an even greater bottleneck than ever.**

In what follows, I will argue that: **Accelerating science calls for foundational advances in Artificial Intelligence and the translation of the resulting advances into cognitive tools that amplify, augment, and extend human intellect and abilities through advanced cyberinfrastructure for science.**

Algorithmic Abstractions of Scientific Domains. Today, algorithmic abstractions increasingly play, across many sciences, the role played by calculus or more generally, mathematics, in the emergence of physics. For example, in biology, we will have a theory of protein folding when we can specify an algorithm that takes as input, a linear sequence of amino acids that make up the protein (and the relevant features of the cellular environment in which folding is to occur), and produces as output, a description of the 3-dimensional structure of the protein (or more precisely, a set of stable configurations). In cognitive science, we will have a theory of learning from experience, when we have an algorithm that learns from observations and experiments. Algorithmic abstractions of the relevant natural entities, relations, and processes in a scientific domain (e.g., biology) allow us to examine the domain through the computational lens and formulate and answer scientific questions in the domain in algorithmic terms. Once created, the algorithmic abstractions become first class computational artifacts in their own right that can be

analyzed, shared, and integrated with other related artifacts, contributing to the acceleration of science. **The creation of sufficiently expressive, yet practically useful algorithmic abstractions of scientific domains calls for major advances in AI, including in particular, knowledge representation, knowledge elicitation, and machine learning. Cyberinfrastructure for science must provide the necessary tools and infrastructure for the collaborative creation, sharing, and use of algorithmic abstractions of scientific domains.**

Algorithmic Abstractions of the Scientific Process: The scientific enterprise (See Figure 1), entails acquiring, organizing, verifying, validating, integrating, analyzing, reasoning with, and communicating information bearing scientific artifacts, namely, experiments, data, models, hypotheses, theories, and explanations associated with natural or built systems lie at the heart of the scientific enterprise. Hence, computing, which offers a powerful medium for digital representation and manipulation of information artifacts offers a powerful formal framework and exploratory apparatus for science. It also offers the theoretical and experimental tools for the study of the feasibility, structure, expression, and, when appropriate, automation of (aspects of) the scientific process, the structure and organization of collaborative teams, modeling the evolution of scientific disciplines, and measuring the impact of scientific discoveries. **The creation and realization of algorithmic abstractions of the scientific process calls for foundational advances across multiple areas of AI, including knowledge representation, planning, optimization, search, multi-agent communication and coordination, natural language processing, information extraction, machine learning, among others. Cyberinfrastructure for science must provide generalizable, modular, extensible, interoperable infrastructure and tools for accelerating science by (i) whenever feasible, automating aspects of science (well beyond building predictive models from data using machine learning, and compute and data intensive simulations).**

Cognitive tools for Amplifying, Augmenting, and Extending Human Intellect and Abilities: Accelerating science requires effective computational tools for mapping the current state of knowledge in a discipline and identifying the major gaps; Generating and prioritizing questions that are ripe for investigation; Extracting and organizing descriptions of experimental protocols, scientific claims, supporting assumptions, and validating scientific claims from scientific literature, and increasingly scientific databases and knowledge bases; Literature-based discovery, including methods for drawing inferences and generating hypotheses from existing knowledge in the literature (augmented with discipline-specific databases and knowledge bases of varying quality when appropriate), and ranking the resulting hypotheses; Expressing, reasoning with, and updating scientific arguments (along with supporting assumptions, facts, observations), including languages and inference techniques for managing multiple, often conflicting arguments, assessing the plausibility of arguments, their uncertainty and provenance; Observing and experimenting, including describing and harmonizing the measurement process and data models, capturing and managing data provenance, describing, quantifying the utility, cost, and feasibility of experiments, comparing alternative experiments, and choosing optimal experiments (in a given context);Navigating the spaces of hypotheses, conjectures, theories, and the supporting observations and experiments; Analyzing and interpreting the results of observations and

experiments, including modeling the measurement process, its bias, noise, resolution; incorporating constraints e.g., those derived from physics, into data- driven inference; closing the gap between model builders and model users by producing models that are expressible in representations familiar to the disciplinary scientists; Synthesizing, in a principled manner, the findings, e.g., causal relationships from disparate experimental and observational studies. **The development of cognitive tools for scientists requires foundational advances in AI. Realizing the promise and potential of such cognitive tools to accelerate science requires advanced cyberinfrastructure for implementing, validating, deploying, and operating the tools.**

Advanced Cyberinfrastructure for Collaborative Science: Because major activities in science increasingly require on collaboration across disciplinary as well as organizational boundaries, there is a need for data and computational infrastructure to support: Organizing and participating in team projects, including tools for decomposing tasks, assigning tasks, integrating results, incentivizing participants, and engaging large numbers of participants with varying levels of expertise and ability in the process; Collaborating, communicating, and forming teams with partners with complementary knowledge, skills, expertise, and perspectives on problems of common interest (including problems that span disciplinary boundaries or levels of abstraction, and call for collaboration across government, industry and academia); Creating and sharing of human understandable and computable representations of the relevant artifacts, including data, experiments, hypotheses, conjectures, models, theories, workflows, etc. across organizational and disciplinary boundaries; Documenting, sharing, reviewing, replicating, and communicating entire studies in the form of reproducible and extensible workflows (with provision for capturing data provenance); Automating the discovery, adaptation, and when needed, assembly of complex analytic workflows from available components; Communicating results of studies or investigations and integrating the results into the larger body of knowledge within or across disciplines or communities of practice; Tracking scientific progress, the evolution of scientific disciplines, and impact on science, engineering, or public policy. **Amplifying, augmenting, and extending human intellect and abilities to accelerate science, calls for fundamental advances in AI, especially, human-machine, machine-machine, and machine mediated human-human collaboration; and advanced cyberinfrastructure collaborative science across disciplinary and institutional boundaries.**

Transparent, trustworthy, accountable cyberinfrastructure for science: As scientific advances rely on data (including sensitive data), infrastructure, and tools beyond those that any individual scientist can fully comprehend or manage, there is a need for: Computable data access and usage agreements that can be enforced within a secure cyberinfrastructure; Audit mechanisms that can be used to verify compliance with the applicable data access and use policies; Repositories of data and their usage agreements that can be adapted and reused in a variety of settings; Agile and secure computing and network services and protocols that can accommodate different types and vintages of instruments; Access privileges that are responsive to the changing needs and roles of individuals; Distributed data management systems or virtual federated collaboratories that enable seamless sharing of data, computational resources, analysis tools, and results across disciplinary and organizational boundaries while ensuring compliance with applicable security as

well as data access and use policies; Sustainable model for data and long-term preservation of both data and the software needed to make use of it; Data and software provenance and other mechanisms for ensuring transparency and reproducibility of data analysis, modeling, etc., detecting, and correcting for implicit or explicit biases or errors in the data as well as the algorithms. **Accelerating science calls for advanced cyberinfrastructure, policy frameworks, and tools for ensuring the transparency, trustworthiness, and accountability of advanced cyberinfrastructure for science.**

Education and Training: Realizing the promise and potential of advances in AI and cyberinfrastructure to accelerate science calls for: a diverse cadre of scientists who combine deep expertise in a scientific domain that have the knowledge and skills to develop and utilize algorithmic abstractions within their scientific domain; interdisciplinary teams of scientists and engineers to design, implement, and study end-to-end systems that flexibly integrate the relevant cognitive tools into complex workflows to solve broad classes of problems in specific domains; organizational, social, behavioral and cognitive scientists to study cyberinfrastructure enabled team science and discover, and translate to practice how best to organize and incentivize such teams to optimize their effectiveness; and organizational changes and funding models that catalyze the acceleration of science through advances in AI and cyberinfrastructure

Understanding the Effect of Data Quality on Analysis Results

Bhowmick, Sanjukta

The success of artificial intelligence(AI) and machine learning (ML) techniques is dependent on the quantity and quality of data. Despite the promise of data deluge, in reality, it is very difficult to obtain large amounts of representative data. This is due to many factors including, (i) the cost of gathering data, (ii) inconsistencies and bias in reported data and (iii) privacy issues dealing with data sharing. The technological, social and legal constraints of data gathering are unlikely to change soon. Thus, a crucial challenge for AI/ML researchers is to determine the how to deliver accurate results in face of these constraints.

Although this problem is prevalent in all areas of AI/ML, I will specifically discuss in the context of networks analysis. Networks (or graphs) are mathematical models of complex systems of interacting entities. Analysis of the network structure provides insights to the properties of the underlying system. Network analysis is used in many disciplines from identifying genes with similar functions in gene correlation networks to recommendation systems in online markets. We consider issues with respect to these three types of poor-quality data;

Incomplete Data. This is data where parts of the information, here edges or vertices are missing. The challenge lies in either filling in the missing information—such as through link prediction or in evaluating by how much the missing data affects the results. Some of our preliminary results show that the *accuracy is determined by not so much as the number of edges missing, but by the position of edges in the network*.

Erroneous/Biased Data. Erroneous data is distinguished from incomplete data, in that spurious edges/vertices that were not in the original network may have been added. Erroneous data can be detected by gathering the data multiple times, and comparing the associated network models. However, this is an extremely time consuming process, and often not feasible.

An analytical approach would be to *test the sensitivity of the data under perturbation*. Under appropriate perturbation models, an unbiased data should give approximately the same quality of results, whereas biased data would clearly show different results under perturbation

Anonymized Data. Due to privacy concerns, data is often shared in an anonymized form. A typical form of anonymization is to ensure that each node in the network shares some neighbors with k-other nodes. An important would be to study how this *change affects the network analysis results, and to design anonymization techniques that can adapt* to the type of analysis being performed, without sacrificing the security.

To summarize, the quality and correctness of data, is an important concern for all learning algorithms. To date, there are not any standardized measures on how to evaluate quality of data or how to measure the sensitivity of results with respect to change in data. Moreover, most anonymization techniques are very general and not tuned to the analysis needs. These questions are critical to understand the limits of AI when developing smart cyberinfrastructure and to design techniques to improve the data quality.

Cyberinfrastructure for science-driven analytics with limited data

Pavan Turaga
Arizona State University
Associate Professor in Electrical Engineering and Media Arts

Sensing modalities have been exploding in areas as diverse as materials science, biology, geographical and space science, with the resultant data as diverse as images, time-series, point-clouds, and functional-data. However, with the rise of such scientific datasets, analytics techniques to convert the raw data to actionable insights have lagged behind. This is because in many of these areas, it is time-consuming to provide detailed human annotation of key concepts, or labels, that can be used to train effective machine learning techniques. Further, in these applications, often the goal is not to perform simple tasks like classification, but to arrive at scientific insights. Due to these reasons, standard applications of neural networks to these domains has resulted in slower progress than anticipated.

For instance, in material science imaging data, for use in 4D material characterization, the imaged data is known to have characteristics at different spatial scales, very different from natural images. This makes standard application of techniques like image segmentation and classification difficult. Further, annotating materials datasets is a highly time-consuming process and acquiring new samples are also time consuming. Simple data augmentation methods are currently used to make visual analytics with deep-nets more robust but also fail as augmentation methods are limited to simple effects like rotations, affine transforms, noise, blur and other factors; increasing the training-set requirements and training-time. Their generalizability beyond the factors considered is often unknown in scientific analytics.

To overcome these limitations, new deep-learning architectures motivated by directly encoding physics-based constraints may help. These constraints include knowledge of imaging physics, illumination models, view-invariant representations, and invariance to image-quality degradation. Turaga work at the Geometric Media Lab uses methods rooted in geometry and topology to enforce these constraints analytically either in loss functions, or in constraints over latent spaces. This can lead to robust architectures providing higher performance under new imaging modalities especially under low-shot learning paradigms.

We believe that we are at an important juncture where interdisciplinary insights from scientific domains, machine learning, cyberinfrastructure can come together to develop a new class of flexible techniques that are:

- rooted in known domain science
- leverage new mathematics from geometry, topology, functional-analysis
- leverage existing cyberinfrastructure from other domains
- adaptable to small datasets
- provide robust, interpretable, and actionable scientific insights.

Geometric Media Lab

NSF Smart Cyberinfrastructure Workshop White Paper:
William Tang, Princeton University/PPPL

Title: "Features of a Possible NSF Smart Cyberinfrastructure Roadmap Supporting Science Applications"

Focus: Associated needs & requirements with illustrative exemplar application from Fusion Energy

---

The basic motivation behind NSF's strong current interest in developing a "Smart Cyberinfrastructure (CI) Roadmap" is to appropriately address the challenge of integrating emerging new smart technologies in Artificial Intelligence/Deep Learning/Machine Learning into existing and future scientific investigations involving the synergistic engagement of the Agency's CI and applications communities. This process can begin by moving toward a common understanding of the current and evolving requirements of "Smart CI" in terms of architectures, algorithms, best practices, enabling technologies, and current gaps in CI for the present and near future,

While there is abundant evidence of great national and international enthusiasm for launching major AI-centric new R&D programs, a logical starting point is to address the following basic set of 6 associated questions: (1) What exciting grand challenge problems can best be addressed with AI-enabled cross-disciplinary cyberinfrastructure" ?; (2) What transformational ideas for associated efficient coupling of simulations, experiments, & computing facilities are needed to accelerate progress in scientific discovery?; (3) What kinds of modern computer architectures and associated infrastructure can best serve the needs of AI-enabled discovery science ?; (4) Since the coupling of AI and HPC is a well-recognized to be a huge opportunity for AI-enabled cyberinfrastructure, what approaches should be adopted ? (6) How can we best enhance exciting opportunities for international and industrial collaborations in an AI-enabled ecosystem?

We can begin by considering some guiding principles for possible NSF Cross-Disciplinary AI Institutes with key themes that include, for example: (1) Trustworthy AI; (2) Foundations of Machine Learning; and (3) AI for Discovery in Physics. The overall, targeted work-scope should cover both foundational as well as "use-inspired" AI research to be explored. Another key element of such interdisciplinary computational institutes would be to further strengthen NSF's mission for education & training of current and future generations of the US workforce -- with the skills to develop and apply AI tools and technologies with innovative impact on today's economy and jobs of the future. This would of cource focus on AI-enabled approaches for education and development of the Nation's graduate and undergraduate students, post-doctoral researchers, and skilled technical workforce. Associated NSF Cross Disciplinary AI institutes could involve universities as well as NSF supercomputing centers (e.g., SDSC, TACC, NCSA, …) who would be well positioned to conduct courses in Deep Learning/Machine Learning with associated connections to workshops and "hackathons" for enabling timely assimilation of new concepts and methodologies in AI-enabled discovery science. The institutes should also be encouraged to engage and establish attractive connections to leading industries – such as Microsoft, Nvidia, Intel, Google-Brain, & Facebook -- to help inject an exciting level of practical connections to accelerated deployment of modern technology.

It is important to create a sense of excitement that requires identifying scientific applications that motivate/energize creation of such a new NSF Smart Cyberinfrastructure. For example -- CNN's "MOONSHOTS for 21st CENTURY" (Hosted by Fareed Zakaria) – has identified 6 grand challenge areas that are easily understandable to the general public to be hisworthy enterprises to be targeted. Fusion Energy is a compelling exemplar in this set, and an associated inspirational endorsement from Stephen Hawking came from his BBC Interview, 18 Nov. 2016, where he commented: "I would like nuclear fusion to become a practical power source. It would provide an inexhaustible supply of energy, without pollution or global warming."

At this point, we will now focus for the rest of this White Paper on using Fusion Energy as an exemplar for illustrating associated needs and requirements associated with "Smart Cyberinfrastructure." Here the major grand challenge involves the demonstration of the scientific and technological feasibility of delivering fusion power in the multinational ITER project – with the highest priority being to accurately predict & control disruptions for the $25B burning plasma ITER experiment that has the goal of exceeding "break-even" (or "power in = power out") by a factor of 10 to 20.

<u>What can AI-enabled advanced cyberinfrastructure deliver today in this Exemplar area?</u>

Artificial Intelligence/Deep Learning brings new technology to accelerate progress
"Predicting Disruptive Instabilities in Controlled Fusion Plasmas through Deep Learning"
NATURE: April, 2019 : Princeton's Fusion Recurrent Neural Network code (FRNN) uses convolutional & recurrent neural network components to integrate both spatial and temporal information for predicting disruptions in tokamak plasmas with unprecedented accuracy and speed on top supercomputers worldwide.

Things we can do in FES with AI now & on near term horizon in this Exemplar area include:
i) Learn predictive models from data without relying upon analytic theory or deep mechanistic understanding: Example: predicting dangerous disruptive events in tokamaks using AI/Deep Learning on huge measured data base enabled by training on leadership class supercomputers achieving unprecedented accuracy and speed.
ii) Initiate moving from prediction to active real-time plasma control:
Example: Introduce a "software integration wrapper" enabling conversion of deep learning predictors written in modern Python/Keras language into conventional plasma control systems written in the older C language.
iii) Develop new software to help explain reasons for deep learning predictive accuracy:
Example: Introduce capability to output not only the "disruption score" for the probability of a disruption event but also a "sensitivity score" in real-time to indicate the underlying physics reasons for the Imminent disruption. ➔ physics-based interpretability + targeted guidance for control actuators upon Implementation into a modern plasma control system (PCS).
Iv) Carry out convergence studies of HPC advances with DL/AI predictive workflow:
Example: First obtain realistic pre-disruption classifiers (e.g. for experimentally-observed "neoclassical tearing modes) from "reduced models" derived from 1st-principles-based electromagnetic PIC simulation results carried out on the SUMMIT supercomputer using exascale class GTC code – and then insert into the AI/DL workflow.

<u>What can AI-enabled advanced cyberinfrastructure deliver within 5 years in this Exemplar area?.</u>

• Successful Development of:
i) Efficient & realistic control strategies based on advanced AI/DL predictors emerge for optimization of performance and avoidance of disruptions in initial operations of ITER;
ii) Continuous vetting of stable, scalable, portable control systems and associated methodology on existing tokamaks (e.g., DIII-D, JET, KSTAR, … leading to the large superconducting JT60-SA tokamak in Japan and finally to ITER)
Cross-disciplinary Exploration:
iii) Novel real-time control methods learned from expertise residing in application areas such as robotics and self-driving cars;
→ Enables leveraging vast experience from well established institutions, including (e.g., Alan Turing Institute in UK); active industrial engagement, such as Microsoft, NVIDIA, INTEL, GOOGLE, FACEBOOK, …..)
• Automation & Acceleration of Discovery Science:
• Systematically moving from (i) New Planning, to Creative Conjecture, to (ii) Further

Experimentation, to (iii) Confirmation/Validation and Integration of New Analysis
➔ Vision of "End-to-End" automated harvesting of real-time scientific insights from operating modern experimental facilities

What can AI-enabled advanced cyberinfrastructure deliver after 10 years in this Exemplar area?

• Realistic real-time models to reduce actual numbers of necessary experiments as the experiments incorporate "lessons learned" from such AI-enabled higher physics fidelity models
• AI becomes common part of scientific laboratory activities.
• AI infuses new scientific, engineering, and operations methodologies into FES
➔ Development of "Virtual Experiments" emerging from validation vs. large trustworthy data bases with associated sensitivity testing to provide an essential filter against "over-hyped" claims
• Improved theoretical formulations from integration of new AI/DL-enabled statistically-validated insights to long-standing plasma physics/FES theories remove/reduce problematic/uncertainty areas
• Validated theory becomes data source for next-generation AI in FES, so that AI begins to contribute to advancing fundamental Plasma Physics/FES theory
• AI-enabled in-depth pursuit of creative FES physics scenarios – e.g., advanced materials, such as, high-temperature superconducting large magnetic field to enable modularity and compactness.

Concluding Comments/Observations on Impact of AI-enabled Advanced Cyberinfrastructure:

• The rapid growth of AI/DL/AI in prominent application domains are likely to mirror the revolutionary trends seen in the business world today – e.g., the re-formation of Amazon and other top businesses that have incorporated AI/DL/ML
https://www.wired.com/story/amazon-artificial-intelligence-flywheel/

• Cross-disciplinary engagement can yield accelerated mutually beneficial results (e.g., Cancer & Clean Energy Fusion R&D)

- Cancer Research → Reference: "Candle Project" with ECP Exascale Computing Project (DOE & NIH) to identify optimal cancer treatment strategies, by building a scalable deep neural network code called the CANcer Distributed Learning Environment (CANDLE).

→ development of predictive models for drug response, and automation of the analysis of information from millions of cancer patient records -- via developing, implementing, & testing DL/AI Algorithms and their benchmarks – such as hyperparameter tuning and associated complex workflows

→ development of predictive models (as just illustrated) in Clean Fusion Energy R&D have followed similar AI/DL approaches

➔DL/AI advances in very different key application Grand Challenge areas like Clean Energy Fusion and Cancer Research can stimulate enhanced cross-disciplinary efforts to leverage connections to enormous worldwide investments in AI/DL/ML R&D !