AI FOR SCIENCE

RICK STEVENS
VALERIE TAYLOR
Argonne National Laboratory
July 22–23, 2019

JEFF NICHOLS
ARTHUR BARNEY MACCABE
Oak Ridge National Laboratory
August 21–23, 2019

KATHERINE YELICK
DAVID BROWN
Lawrence Berkeley
National Laboratory
September 11–12, 2019

U.S. DEPARTMENT OF ENERGY | U.S. DEPARTMENT OF ENERGY | Office of Science | ASCR

February 2020

Cover credit: Image by Raigvi/Shutterstock; Baivector/Shutterstock

AI for Science

Report on the Department of Energy (DOE) Town Halls on
Artificial Intelligence (AI) for Science

Town Hall Co-Chairs

Rick Stevens — Associate Laboratory Director, Argonne National Laboratory
Jeffrey Nichols — Associate Laboratory Director, Oak Ridge National Laboratory
Katherine Yelick — Associate Laboratory Director, Lawrence Berkeley National Laboratory

Department of Energy Contact

Barbara Helland — Program Manager, Department of Energy

Special Assistance

Chapter Leads:

Argonne National Laboratory
Valerie Taylor, Director, Mathematics and Computer Science Division

Mihai Anitescu, Prasanna Balaprakash, Pete Beckman, Thomas S. Brettin, Charles E. Catlett, Andrew Chien, Santanu Chaudhuri, Ian Foster, Dogan Gursoy, Salman Habib, Cynthia Jenks, Rao Kotamarthi, Zein-Eddine Meziani, Michael E. Papka, Robert Ross, Stefan Wild

Lawrence Berkeley National Laboratory
David Brown, Director, Computational Research Division

Katerina Antypas, Wes Bethel, Ben Brown, Paolo Calafiura, Wibe de Jong, Sudip Dosanjh, Inder Monga, Peter Nugent, Mary Ann Piette, Prabhat, Brian Quiter, Lavanya Ramakrishnan, John Shalf, Haruko Wainwright, John Wu, Petrus Zwart

Oak Ridge National Laboratory
Arthur Barney Maccabe, Director, Computer Science and Mathematics Division

David Dean, James Hack, Kenneth Herwig, Judith Hill, Forrest M. Hoffman, Teja Kuruganti, Bronson Messer, Nageswara Rao, Arjun Shankar, Bobby G. Sumpter, Georgia Tourassi, John Turner, Jeffrey Vetter, David Womble, Steven Young

Lawrence Livermore National Laboratory
Ana Kupresanin

General Atomics
David Humphreys

Administrative:

Argonne National Laboratory: Silvia Mulligan
Lawrence Berkeley National Laboratory: Hellen Cademartori
Oak Ridge National Laboratory: Becky Verastegui

Publication: **Argonne National Laboratory**: Linda Conlin, Kristen Dean, Lorenza Salinas, John W. Schneider, Sonya Soroko

Editorial: **Argonne National Laboratory:** Emily M. Dietrich, Laura Wolf
Lawrence Berkeley National Laboratory: Carol Pott
Oak Ridge National Laboratory: Scott Jones, Elizabeth Rosenthal

Contents

Executive Summary **1**

Introduction: AI for Science **5**
Materials, Environmental, and Life Sciences 5
High-Energy, Nuclear, and Plasma Physics 7
Engineering, Instruments, and Infrastructure 9
Foundations, Software, Data Infrastructure, and Hardware 11
Conclusions 14

01. Chemistry, Materials, and Nanoscience **17**
1. State of the Art 17
2. Major (Grand) Challenges 18
3. Advances in the Next Decade 21
4. Accelerating Development 23
5. Expected Outcomes 25
6. References 25

02. Earth and Environmental Sciences **27**
1. State of the Art 27
2. Major (Grand) Challenges 28
3. Advances in the Next Decade 31
4. Accelerating Development 31
5. Expected Outcomes 34
6. References 34

03. Biology and Life Sciences **37**
1. State of the Art 37
2. Major (Grand) Challenges 38
3. Advances in the Next Decade 41
4. Accelerating Development 42
5. Expected Outcomes 43
6. References 43

04. High Energy Physics **45**
1. State of the Art 46
2. Major (Grand) Challenges 47
3. Advances in the Next Decade 50
4. Accelerating Development 51
5. Expected Outcomes 52
6. References 52

05. Nuclear Physics **55**
1. State of the Art 56
2. Major (Grand) Challenges 58
3. Advances in the Next Decade 61
4. Accelerating Development 62
5. Expected Outcomes 63
6. References 63

Contents (Cont.)

06. Fusion 65
1. State of the Art 65
2. Major (Grand) Challenges 66
3. Advances in the Next Decade 69
4. Accelerating Development 70
5. Expected Outcomes 70
6. References 71

07. Engineering and Manufacturing 73
1. State of the Art 73
2. Major (Grand) Challenges 74
3. Advances in the Next Decade 77
4. Accelerating Development 78
6. References 79

08. Smart Energy Infrastructure 81
1. State of the Art 81
2. Major (Grand) Challenges 83
3. Advances in the Next Decade 85
4. Accelerating Development 86
5. Expected Outcomes 87
6. References 87

09. AI for Computer Science 89
1. State of the Art 89
2. Major (Grand) Challenges 92
3. Advances in the Next Decade 94
4. Accelerating Development 95
5. Expected Outcomes 96
6. References 96

10. AI Foundations and Open Problems 99
1. State of the Art 99
2. Major (Grand) Challenges 100
3. Advances in the Next Decade 102
4. Accelerating Development 105
5. Expected Outcomes 105
6. References 106

11. Software Environments and Software Research 109
1. State of the Art 109
2. Major (Grand) Challenges 109
3. Advances in the Next Decade 113
4. Accelerating Development 114
5. Expected Outcomes 114
6. References 115

Contents (Cont.)

12. Data Life Cycle and Infrastructure 117
1. State of the Art 118
2. Major (Grand) Challenges 119
3. Advances in Next Decade 122
4. Accelerating Development 123
5. Expected Outcomes 123
6. References 123

13. Hardware Architectures 125
1. State of the Art 125
2. Major (Grand) Challenges 126
3. Advances in the Next Decade 129
4. Accelerating Development 129
5. Expected Outcomes 131
6. References 131

14. AI for Imaging 133
1. State of the Art 133
2. Major (Grand) Challenges 135
3. Advances in the Next Decade 136
4. Accelerating Development 137
5. Expected Outcomes 138
6. References 138

15. AI at the Edge 141
1. State of the Art 142
2. Major (Grand) Challenges 143
3. Advances in the Next Decade 145
4. Accelerating Development 146
5. Expected Outcomes 147
6. References 147

16. Facilities Integration and AI Ecosystem 149
1. State of the Art 149
2. Major (Grand) Challenges 149
3. Advances in the Next Decade 152
4. Accelerating Development 153
5. Expected Outcomes 153

AA. Report Writing Team 155

AB. Agendas 157
Argonne National Laboratory 157
Oak Ridge National Laboratory 160
Lawrence Berkeley National Laboratory 163
Washington, DC 167

Contents (Cont.)

AC. Combined Town Hall Registrants 171

AD. Abbreviations and Terminology 197

AE. References 201

Executive Summary

From July to October 2019, the Argonne, Oak Ridge, and Berkeley National Laboratories hosted a series of four town hall meetings attended by more than 1,000 U.S. scientists and engineers. The goal of the town hall series was to examine scientific opportunities in the areas of artificial intelligence (AI), Big Data, and high-performance computing (HPC) in the next decade, and to capture the big ideas, grand challenges, and next steps to realizing these opportunities.

In this report and in the Department of Energy (DOE) laboratory community, we use the term "AI for Science" to broadly represent the next generation of methods and scientific opportunities in computing, including the development and application of AI methods (e.g., machine learning, deep learning, statistical methods, data analytics, automated control, and related areas) to build models from data and to use these models alone or in conjunction with simulation and scalable computing to advance scientific research.

The AI for Science town hall discussions focused on capturing the transformational uses of AI that employ HPC and/or data analysis, leveraging data sets from HPC simulations or instruments and user facilities, and addressing scientific challenges unique to DOE user facilities and the agency's wide-ranging fundamental and applied science enterprise.

The town halls engaged diverse science and user facility communities, with both discipline- and infrastructure-specific representation. The discussions, captured in the 16 chapters of this report, contain common arcs revealing classes of opportunities to develop and exploit AI techniques and methods to improve not only the efficacy and efficiency of science but also the operation and optimization of scientific infrastructure.

The community's experience with machine learning (ML), HPC simulation, data analysis methods, and the consideration of long-term science objectives revealed a growing collection of unique and novel opportunities for breakthrough science, unforeseeable discoveries, and more powerful methods that will accelerate science and its application to benefit the nation and, ultimately, the world.

New AI techniques will be indispensable to supporting the continued growth and expansion of DOE science infrastructure from ESnet to new light sources to exascale systems, where system scale and complexity demand AI-assisted design, operation, and optimization. Toward this end, novel AI approaches to experiment design, *in-situ* analysis of intermediate results, experiment steering, and instrument control systems will be required.

DOE's co-design culture involving teams of scientific users, instrument providers, mathematicians and computer scientists can be leveraged to develop new capabilities and tools such that they can be readily applied across the agency's (and indeed the nation's) diversity of instruments, facilities, and infrastructure. This report captures some early opportunities in this direction, but much more needs to be explored.

From chemistry to materials sciences to biology, the use of ML and deep learning (DL) techniques opens the potential to move beyond today's heuristics-based experimental design and discovery to AI-enhanced strategies of the future.

Early use of generative models in materials exploration suggests that millions of possible materials could be identified with desired properties and functions and evaluated with respect to synthesizability. The synthesis and testing stages necessary for such scales will in turn rely on ML and adaptive, autonomous robotic control of high-throughput synthesis and testing lines, creating "self-driving" laboratories.

The same complexity challenge and concomitant need to move from human-in-the-loop to AI-driven design, discovery, and evaluation also manifests across the design of scientific workflows, optimization of large-scale simulation codes, and operation of next generation instruments.

Exascale systems and new scientific instruments, such as upgraded light sources and accelerators, are increasing the velocity of data beyond the capabilities of existing instrument data transmission and storage technologies. Consequently, real-time hardware is needed to detect events and anomalies in order to reduce the raw instrument data rates to manageable levels. New ML, including DL, capabilities will be critically important in order to fully exploit these instruments, replacing pre-programmed hardware event triggers with algorithms that can learn and adapt, as well as discover unforeseen or rare phenomena that would otherwise be lost in compression.

In recent years, the success of DL models has resulted in enormous computational workloads for training AI models, representing a new genre of HPC resource demand. Here, the use of AI techniques to optimize learning algorithms and implementation will be necessary with respect to both the energy cost of large-scale computation and to the exploitation of new computing hardware architectures. AI in HPC has already taken the form of neural networks trained as surrogates to computational functions (or even entire simulations), demonstrating the potential for AI to provide non-linear improvements of multiple orders of magnitude in time-to-solution for HPC applications (and, coincidentally, reductions in their cost).

Similarly, scientific infrastructure—accelerators, light sources, networks, computation and data resources—have reached scales and complexities that require the use of ML for tasks such as anomaly detection in operational data (e.g., for cybersecurity). Moving from today's fixed rules-based operating procedures to the use of AI algorithms that factor real-time analysis will be indispensable for optimizing performance and energy use of increasingly complex, large-scale infrastructures. New DL methods are required to detect anomalies and optimize operating parameters, with additional potential to predict failures as well as to discover new optimization algorithms and novel mechanical or externally induced threats.

The DOE computing facilities such as Summit, Perlmutter, Aurora and Frontier will simultaneously support the development of existing large-scale simulations, new hybrid HPC models with AI surrogates, and the exploration of new types of generative models emerging from multimodel data streams and sources. Future systems envisioned over the next decade may need to support even richer workloads of traditional HPC and next-generation AI-driven scientific models.

AI will not magically address these and the other opportunities and challenges discussed in this report. Much work will be required within all science disciplines, across science infrastructure, and in the theory, methods, software, and hardware that underpin AI methods. The use of AI to design and tune hardware systems—whether exascale workflows, national networks, or smart energy infrastructure—will require the development and evaluation of a new generation of AI frameworks and tools that can serve as building blocks that can be adapted and reused across disciplines and

across heterogeneous infrastructure. Bringing AI to any specific domain—whether it is nuclear physics or biology and life sciences—will demand significant effort to incorporate domain knowledge into AI systems, quantify uncertainty, error, and precision, and appropriately integrate these new mechanisms into state-of-the-art computational and laboratory systems.

The overflowing attendance at the AI for Science town halls, the level of enthusiasm and the engagement of attendees, the number of spontaneous AI projects throughout every scientific discipline, and the commitment to growth in this area at the nation's premiere laboratories all combine to indicate that the DOE scientific community is ready to explore and further the transformational potential of AI through 2030 and beyond.

This page intentionally blank.

Introduction: AI for Science

The AI for Science town halls brought together more than a thousand researchers from DOE National Laboratories, industry, and academia to identify opportunities for AI to impact the national science enterprise supported by DOE. The teams also outlined the research and infrastructure needed to advance AI methods and techniques for science applications.

Sixteen topical expert teams summarized the state of the art, outlined challenges, developed an AI roadmap for the coming decade, and explored opportunities for accelerating progress on that roadmap.

Important themes emerged for AI applications in science. For example, participants anticipate the use of AI methods to accelerate the design, discovery, and evaluation of new materials, and to advance the development of new hardware and software systems; to identify new science and theories within increasingly high-bandwidth instrument data streams; to improve experiments by inserting inference capabilities in control and analysis loops; and to enable the design, evaluation, autonomous operation, and optimization of complex systems from light sources to HPC data centers; and to advance the development of self-driving laboratories and scientific workflows.

Important themes also emerged with respect to outlining the research needed to advance AI. For example, participants highlighted the need to incorporate domain knowledge into AI methods to improve the quality and interpretability of the models; the need to develop software environments to enable AI capabilities to seamlessly integrate with large-scale HPC models; and the need to automate the large-scale creation of "FAIR" (findable, accessible, interoperable, and reusable) data, given the central role of data in an AI-centric future science landscape.

Below, we briefly outline the principle findings of the main sections of the report.

Materials, Environmental, and Life Sciences

Chapters 1–3

Finding new materials, chemical compounds, and biological agents able to address contemporary challenges—for example, batteries with 10 times more storage capacity, materials that capture more solar energy at greater efficiency, and new drugs targeting emerging pathogens—is a grand challenge due to the nearly infinite chemical, biological, and atomic design spaces to which scientists have access. Such discovery requires pervasive AI-enabled automation, from experiment design to execution and analysis.

Projecting environmental risk and developing resiliency in a changing environment are central challenges to earth and environmental sciences, encompassing atmosphere, land, and subsurface systems along with their interdependencies. From large-scale observatories such as the Atmospheric Radiation Measurement (ARM) facility, AI methods will be essential to obtaining the data needed to refine complex earth and environmental systems models, and to developing new models with unprecedented fidelity and resolution. AI "at the edge"—where people and things meet—will enable autonomous observatories to detect anomalies and outliers, adapting instrument settings and algorithms to provide detailed measurement of events and conditions that would otherwise go unnoticed.

Biology and life sciences are at the vanguard of AI applications, for instance using population genomics data to learn the bases of complex traits and discovering or building workflows that automate the inverse design of microbial and plant cells. "Self-driving" laboratories will

leverage new generative models and reinforcement learning to explore potential compounds for cancer drugs, evaluate their synthesizability, or model their response in target tumors.

Discovery and Data

Scientists have used computational approaches to explore virtually materials and chemical compounds, leveraging new data sources containing the simulated properties of millions of simple materials and chemical compounds. Deep learning approaches are being developed to explore more deeply inside vast molecular and biological design spaces. Molecular scientists are using AI to learn force fields to enable near-exact molecular dynamic (MD) simulations with fully quantized electrons and nuclei. Such analyses, intractable only a few years ago, must now be captured and advanced in the form of AI software toolkits and services.

Across the sciences, rapidly growing data sources can, in principle, be used to train ML models provided that the data can be "found, accessed, and are interoperable and reusable," or "FAIR." The use of DL and unsupervised learning for automatic labeling and reduction of data also needs to be captured as adaptable software services that can be applied to data sources ranging from environmental datasets at broad spatial and time scales, to instrument data from materials testing, to genomics data.

For life sciences, energy infrastructure sciences, and even national security, access is needed to protected sensitive data. We must establish new infrastructure to enable shared use of data that cannot be moved or revealed due to privacy concerns. Similar challenges arise with respect to proprietary manufacturing, mobility, and private energy data.

Learning and Integrating Domain Knowledge

Today's computational learning frameworks are not yet able to realize the full potential of AI-enabled materials, chemical, environmental, and biological sciences. We need new AI methods that can both predict complex phenomena and provide insights into underlying processes. Such methods will be foundational to our capacity to design custom biological systems capable of addressing major global health and environmental challenges—that is, ultimately to "build life to spec." Here, as with materials design, AI-enabled, self-driving laboratories (through new automation and decision support services) can fuel game-changing advances in the understanding and deployment of biological, chemical, and environmental systems.

Self-Driving and Steering Laboratories

The most exciting discovery possibilities for emerging instruments such as for bio- or materials imaging lie in going beyond today's human-in-the-loop experimentation, and allowing embedded AI to evaluate results and steer experiments.

AI-assisted management and control of research labs, instruments, facilities, experiments, and workflows can help achieve a variety of goals, for instance by adapting workflows in response to new hypotheses generated during workflow execution, scheduling resources for more efficient use of facility hardware, and dramatically reducing the total cost of operating facilities.

Experimental science is moving rapidly toward more frequent online analysis and adaptation. In "self-driving" laboratories, AI can be used not only for analysis and hypothesis generation, but also to act on intermediate results, adapting

to new data by adjusting experimental parameters or laboratory processes toward specific goals, such as protecting resources, maximizing the data gathered related to a specific phenomenon, or following up on surprising or anomalous results.

AI-guided self-driving laboratories are envisioned that can automate the design, synthesis, and evaluation of material and increase the pace of discovery by orders of magnitude.

AI in HPC

Multi-scale models are needed to understand the underlying systems affecting phenomena associated with the growing global demand for fuel, food, water, and predictable weather. AI technologies can reveal the emergent controls of these enormously complex environmental, plant, and microbial biosystems, enabling us to engineer our environment, for instance to expand the range of arable lands while improving water availability and quality. In order to enable such discovery capabilities, we must not only improve the performance and quality of HPC models (e.g., using ML surrogates) but we must make it possible to build generative models from diverse observations (e.g., time series measurements) and computational simulations. This will need to be aligned with AI-based inverse problem solvers, such as for image-to-phase or waveform-to-source problems to explore novel geoengineered solutions.

Such simulation models represent another domain where AI is already showing transformative results. Time-to-solution of modeling systems and associated reduction in computational needs (and associated energy use) can be improved by combining data-informed AI approximations with physical principles for earth systems, ecosystems, soil microbiology, watershed, and other models. The use of such AI "surrogate" functions will require robust, explainable AI methods for training and validating hybrid models, and the integration of uncertainty quantification into AI workflows.

A secure environment for objective benchmarking of AI algorithms against community consensus metrics is needed to detect, monitor, and possibly correct dataset biases or inconsistent AI performance. Foundational technologies are needed to promote a rigorous statistical framework to monitor for potential biases or inaccuracies in collected data, and to monitor AI performance to confirm robust performance or identify performance gaps. These topics are detailed in *Foundations, Software, Data Infrastructure, and Hardware* (page 11).

High-Energy, Nuclear, and Plasma Physics

Chapters 4–6

In cosmology, high-energy physics, fusion, and nuclear physics, the next decade will bring new, enormous, and rich data sets from new light sources, accelerators, tokamak facilities, and advanced survey telescopes, unparalleled in depth and resolution at the observed scales. These observations will be combined with exascale-enabled simulations modeling structure formation in unprecedented detail to enable major scientific advances. ML, including DL, techniques will be crucial in the analysis of multi-spectral observational data sets. "AI-in-HPC" approaches to simulation that use fast AI-based surrogates will allow the reconstruction of the history of the universe from the Big Bang until today at unprecedented fidelity, from the largest scales down to our own galaxy.

The multiscale, highly correlated, and high-dimensionality nature of the physics of the nuclear force also leads to a rich set of phenomena in nuclear physics. AI techniques offer the possibility of increased understanding and new discoveries via DL analyses of light source experimental data, especially given recent and planned upgrades and resulting

increased data volumes and rates. Fusion scientists look to AI/ML techniques for breakthroughs ranging from maximizing predictive understanding of fusion plasmas and the burning plasma state to enabling real-time control in long-pulse tokamak experiments, and ultimately AI-in-the-loop plasma prediction and control solutions necessary for sustained, safe, and efficient fusion power plant operation.

Discovery and Data

In coming years, the global high-energy physics community will deploy AI-controlled, city-size scientific instruments (particle accelerators and particle detectors) that produce zettabytes of data. Similarly, high-bandwidth streams will come from new survey telescopes, upgraded light sources, and tokamak experiments. AI-powered hardware will be required to filter detector data in microseconds. AI inference systems trained by data and simulations of detector response will be needed to enable high-precision studies, while unsupervised AI-based searches for anomalies and rare events, indeed even for "New Physics," will open new windows for discovery.

Learning and Integrating Domain Knowledge

AI methods are critically important if we are to fully exploit data from new or upgraded large-scale instruments and complex experiments—facilitating the collection, evaluation, and analysis of metadata; improving data reduction and documentation of experimental conditions; and facilitating data interoperability.

To achieve such capabilities across diverse instruments, we must create usable tools for the large-scale training and optimization of ML models, training methodologies that can detect rare features in high-dimensional spaces, and tools to quantify the impact of systematic effects of the accuracy and stability of complex ML models. However, one of the obstacles to applying data science to hypothesis generation and experimental design is the availability (to the general community) and the lack of uniformity of data. A significant need in the coming decade will be to develop ML methods to automatically annotate and structure data from computational models and experimental facilities such as the international ITER Tokamak, upgraded light sources such as the Advanced Photon Source (APS), and Advanced Light Source (ALS).

Designing and Steering Experiments

The introduction of ML and AI into the scientific process for hypothesis generation and the design of experiments promise to significantly accelerate the scientific process by automating and accelerating the development of models and the testing of hypotheses. For this to become reality, domain knowledge must be integrated into ML models, moving beyond current models that are either purely data-driven or that incorporate only simple algorithms, laws, and constraints. ML techniques that combine theoretical and data-driven models in hybrid systems that better represent the underlying dynamics specific to phenomena will be especially key.

Across experimental sciences, AI-aware experimental design, construction, and operation of scientific instruments offer transformative improvements. For detectors and accelerators, the use of reinforcement learning (RL) will both reduce beam generation times and improve the quality of beams delivered to end stations. Improving particle tracking will also rely on ML techniques, but these techniques must be sufficiently validated to ensure the tracking performs on data in the energy region of interest. AI-centric workflows using deep neural networks (DNNs) trained by detector signals will improve our ability to distinguish event candidates from background data.

AI algorithms have demonstrated powerful anomaly detection capabilities and will also provide the necessary performance for

intelligent instrument operation and experiment-steering. ML inference with microsecond latency will be required to support particle physics trigger applications in large detectors and associated event processing operations.

The use of AI for real-time experiment-steering will increasingly become indispensable, whether for light source instruments or tokamak experiments, and will become equally critical for orchestrating the coupling of cosmological models with the steering mechanisms of a new generation of multi-spectral telescopes.

Engineering, Instruments, and Infrastructure

Chapters 7–9 and 14–16

Terms such as “smart manufacturing” and “digital twins” reference transformational approaches for expanding optimization to include an entire manufacturing lifespan, from raw materials to shape/topology to manufacturing process to end use. Concurrently, AI has been used in generative design, a two-step iterative process based on design goals that first generates possible outputs that meet specified constraints and then allows a designer to tune variables to meet constraints. Generative adversarial networks are often used to drive the underlying optimal design.

The nation’s energy infrastructure is moving increasingly from traditional loads (non-digital, invisible) to many more and smaller loads that expose data (are visible) and have communication and intelligence features amenable to a cooperative load-management approach. Combined with increasingly intelligent energy distribution and generation infrastructure, the complexity, nonlinearity, and emergent behaviors of these systems will require AI-enabled, distributed and cooperative configuration, optimization, threat detection and avoidance, and control.

Designing and Steering Infrastructure

Just as AI will enable breakthroughs in automation (such as designing experiments, self-driving laboratories, or steering instruments), it will make it possible for the same techniques to be applied to designing and operating complex infrastructure. From electrical generation to transmission to distribution systems, increasingly powerful sensors—with edge computation enabling AI *in-situ* for anomaly detection, predictive analytics, and controls/optimization—will improve resilience as well as restoration by enabling predictive capabilities of after-event states and sharper awareness during the restoration process. AI-driven, real-time intelligence in this context can perform information fusion from disparate sources, coupling real-time infrastructure data with infrastructure models (e.g., a “digital twin”). Similarly, AI/ML-enabled predictive models trained by infrastructure data will be indispensable for exploring the design spaces for smart energy—as well as transportation—infrastructure, HPC computing systems and data centers, and communications networks.

In similar fashion, particle accelerators, light sources, and complex instruments such as ITER comprise many interconnected subsystems of magnets; mechanical, vacuum, and cooling equipment; power supplies; and other components. These instruments have thousands of control points and require high levels of stability, making their operation a complex optimization problem. The operation of these instruments has benefited from AI/ML-based solutions but remains extremely difficult due to the lack of *a priori* models for reliable and safe control. In the absence of such models, learning models based on raw data and other AI/ML-based solutions have been explored, with promising results.

Even smaller scales, such as manufacturers of limited volume batches of materials and those

that produce many variants of similar designs for customized products, are limited by mere automation with heuristics-based operational rules on robotic assembly lines. As with self-driving laboratories, this widespread class of manufacturing genre must move to robotics with AI-at-the-edge to perform tasks autonomously (in similar fashion, as noted earlier, with respect to self-driving laboratories or remote observatories).

These data-driven methods for control-level modeling, management, and interpretation of real-time data for control, optimal trajectory determination, and real-time prediction to support continuous and asynchronous actions and prevent faults will also accelerate the development of approaches to the operation of new types of infrastructure such as fusion power plants.

DOE also operates instruments with components distributed over distances from hundreds of kilometers (e.g., ESnet or the ARM facility). Moving to autonomy and adaptive measurement makes the current practice of centralized control intractable. Whether in laboratory experiment lines, on city-sized accelerator facilities, or for continental-scale infrastructure, AI will be needed to support infrastructure as autonomous, self-tuning, and self-healing complex systems with emergent properties and non-linear behavior, relying on AI-at-the-edge due to complexity as well as latency and data communications bandwidth.

Commercial AI hardware and system-on-chip (SoC) systems also have a key role to play, given DOE's billions of dollars of investment in experimental facilities. Ultra-low latency and low power inference for scientific experimental control in these facilities can enable more complex, intelligent experiments, and more efficient operation. Again co-design and overall system architecture are critical as even the most time-sensitive commercial applications fueling the AI hardware industry, such as autonomous driving, require millisecond response, while DOE instruments such as electron microscopes and light sources can require responses in the 100 nanosecond range—over 100,000 times faster.

All of DOE's current scientific facilities—ESnet, exascale machines, the continentally distributed ARM, individual light sources, data sources from field-deployed sensors, and instrument and HPC data repositories—have been designed for traditional scientific workflows. Every link in this chain, from data portals and networks to edge systems, HPC resources, and input/output (I/O) systems must evolve to support the new demands of AI applications and workflows.

Infrastructure Security

As critical infrastructures increasingly rely on information systems, AI applications will offer the best approach to detecting and diagnosing cyber and physical attacks and threats in real-time. Removing the human-in-the-loop is increasingly necessary for defensive responses on the same millisecond timescales as digital attacks. Here AI can offer novel techniques, including surrogate models, closure models, and learning-driven compute acceleration of high-fidelity models and solvers.

AI in HPC

As noted above, use of AI surrogates within HPC models has the potential to improve time-to-solution by orders of magnitude, albeit replacing first-principles functions with approximations. AI-based surrogate models can play at least three roles in manufacturing systems, including *a priori* optimization, *in situ* real-time process control, and heterogeneous manufacturing through the transfer of AI models between different devices and/or feedstocks.

With infrastructure and manufacturing, surrogate models could form the basis for digital twins that guide design and operation. Determining the best AI techniques to generate and validate surrogates that are robust and

with minimal bias will be important, along with research to explore, for at least several exemplar manufacturing and infrastructure processes, the optimal type and quantity of data to improve design optimization.

Surrogates must also incorporate an understanding of emergent behaviors of interacting AI agents while capturing the multi-physics of complex infrastructure and energy systems, learning from the combination of measured data and physics- or model-based simulation data for rapid prediction. Critically, new methods for validating and testing AI-based models, controls, and optimization will be required in order to entrust critical services to their control, with verifiable trust being as important as the capabilities themselves.

AI-Driven Leadership Computing

The DOE's Aurora, Frontier, and Perlmutter architectures are already designed to optimize for AI workflows. Follow-on systems at the LCFs and NERSC will have upgrades and enhancements informed by the new AI services, workflows, and toolkits discussed throughout this report. Aligning the future path-forward efforts with the development of these new capabilities will be critical to enabling an AI-based instrument approach to future infrastructure and experiment design.

Instrument-to-Edge

Existing large-scale instruments, upgrades such as those to APS or ALS, and new instruments such as ITER, all share the need for AI services that can exploit their capabilities and plumb the unprecedented volumes of data they produce. The envisioned AI-based services and toolkits as described earlier will have the most impact if undertaken in concert with an "Instrument-to-Edge" hardware and software infrastructure that is developed and incrementally deployed to grow a common control and analysis architecture across DOE's major instruments. Experiments using these instruments will rely on these new AI design, optimization, and control services while also providing data that can use new AI-based services for creating and refining generative models that can guide the optimization and safe operation of the instruments themselves.

Foundations, Software, Data Infrastructure, and Hardware

Chapters 10–13

As noted throughout the disciplinary, engineering, and infrastructure discussions, research and infrastructure are needed to advance AI methods and techniques to address the complex challenges of using AI to advance science discovery. It is recognized that research is needed in areas such as processor and memory design, mathematical AI foundations, software environments, data infrastructure, and hardware.

Training Models

The core of any ML-based AI system is the creation of an abstract model and the training of that model is based on data. Data efficient learning in ML systems must be studied with respect to algorithms and efficiency of implementation, and especially with respect to exploiting new architectures, whether through the use of AI-oriented, reduced precision accelerator hardware or novel computing systems (e.g., quantum, neuromorphic) and associated programming paradigms.

Today's approaches to ML and AI are generally domain-agnostic, ignoring domain knowledge that extends far beyond the raw data itself. For example, current approaches ignore physical laws, available forward simulations, and established invariances and symmetries. Incorporating modeling and simulation capabilities to generate use case specific training data leverages decades of HPC improvements to accelerate learning; incorporating mathematical equations and scientific literature leverages centuries of advances in theory.

New AI Hardware and Systems Components

There is an explosion of new AI hardware in industry, however the target applications driving these devices largely comprise consumer or enterprise areas such as autonomous driving, social networks, e-commerce, and gaming. As evidenced in DOE's Exascale Computing Project, there are significant opportunities to co-design heterogeneous compute nodes that leverage these new architectures and commodity SoC ecosystems.

A set of integrated new AI workflow frameworks and exemplar applications will be needed to evaluate emerging AI architectures from edge SoCs to HPC data centers. This would effectively create both an evaluation tool set and a simultaneous series of specific science-based challenges to drive and shape new AI technologies, including those that fuse explicit knowledge and learned function.

Programming Models and Workflows

The design of next-generation hardware and software systems—from new chips to entire HPC systems—and the mapping of application codes to target systems is currently a static process that involves human-in-the-loop design with repeated experiments, modeling, and design space exploration. As these systems increase in complexity and heterogeneity, current strategies will be impractical.

Early work demonstrating systems and workflows that integrate AI capabilities with traditional HPC simulation has largely involved bespoke capabilities for each experiment. The frameworks, software, and data structures are distinct, and APIs do not exist that would enable even simple coupling of simulation and modeling codes with AI libraries and frameworks. *In situ* data analysis requiring ML capabilities suffers from the same limitations.

To fully realize capabilities ranging from self-driving laboratories to AI-designed, implemented, and operated scientific workflows, new programming and run-time models must also be developed. For example, scientists might ideally describe workflows as high-level goals and itemize building-block tasks (i.e., experiments, simulations) and rough models of the costs of those tasks. An AI system could instead generate a specific workflow, incorporating expert knowledge, to accomplish those tasks, adapting as results are uncovered or new data become available and refining the models of costs (e.g., in energy use or time). Such workflows will need to operate across orders of magnitude variations in communications latency and bandwidth and in computational power and storage, especially in cases of specialized edge devices designed for low-power deployments in the field. These programming frameworks will need to provide resource discovery, matching, negotiation, and complex optimizations of these new forms of heterogeneous distributed computing infrastructure, including the integration of inference on low-power edge systems with iterative learning systems within a few milliseconds of the edge (e.g., in 5G telecommunications stations) and deep learning in data centers.

Current HPC memory and storage systems are architected for traditional HPC simulation-only workloads with relatively small inputs and large outputs, where the access patterns are predictable, contiguous, block-based operations. Current AI training workloads, in contrast, must read large datasets (i.e., petabytes) repeatedly and perhaps non-contiguously for training. AI models will need to be stored and dispatched to inference engines, which may appear as small, frequent, random operations. Indeed, the model for computing within DOE will need to evolve to where specialized AI hardware cooperates with traditional HPC systems to train models that are dispatched to low-power devices at the edge.

AI Foundations

AI presents a unique opportunity for creating data-driven surrogate models that are potentially orders of magnitude faster to run than first-principles simulation codes and that can be particularly effective in the ability to simulate physical processes that span many spatial and temporal scales. Rigorously understanding tradeoffs such as generalization limits, proofs of interpolation/extrapolation, robustness, assessment of confidence associated with predictions, and effects of the input data will impact not only model selection in AI systems, but also the creation and investigation of new classes and types of models.

At the most basic level, frameworks and tools are needed to establish that a given problem is effectively solvable by AI/ML methods and is not subject to limits such as extreme complexity, unbounded problems, or explainability. Principles of theoretical computer science provide a rigorous framework to establish critical properties of AI/ML codes, namely computability, learnability, explainability, and provability.

To become an accepted part of the toolboxes used by scientists and engineers, the validity and robustness of AI techniques need to be trusted. What are the limits of AI techniques, and what assumptions and circumstances can lead to establishing assurance of AI predictions and decisions? Which AI techniques can best address different sampling scenarios and enable efficient AI on various computing and sensing environments? Resulting AI systems must similarly address assurance: whether and when an AI model can be trusted. Why does the AI model work for a problem? What are the internal representations of data that the AI model has learned during training? How can the behavior of the AI model be explained? How confident are the AI models on their predictions given the different sources of uncertainties and inductive biases involved? For such an AI model to be accepted as a well-characterized tool for science, the research community will need to address these questions and develop advanced capabilities to explain the behavior of the AI model.

Especially for systems operating experiments, instruments, or critical infrastructure, validation is vital regardless of whether the AI model is making the right decision for the right reason. Has the AI model learned spurious correlations, or can the model determine the control variables? Can AI be used to identify causal variables or distinguish between cause and effect? Typically this cannot be done with a single training dataset. Instead, the AI model needs to be trained to construct a hypothesis, typically a counterfactual one, and to design an experiment—including the collection of data (and the suitability of that data)—to test that hypothesis.

Opportunities exist for fundamental advances in optimization algorithms, differentiation techniques, and models—foundational to training in AI. Additionally, an important aspect in the development and application of AI is the quantification of uncertainties. Where AI and ML are used in physics-based applications, established approaches to UQ are applicable. In other cases, particularly in classification problems, ML models tend to be highly nonlinear systems that are extremely sensitive to input data, and small (e.g., undetectable to the human eye) changes can lead to misclassification.

Addressing the computer science challenges will require a comprehensive AI/ML science program to develop and refine foundational limits and solvable problems and to sharpen the solutions for solvable classes to ensure effective computation, performance guarantees and explanations. This is an urgent issue, as work on the foundations of AI and ML has been far outpaced by the empirical exploration and use of such techniques—often in the form of bespoke systems with disparate architectures. Consequently, the principles underlying the use and understanding of these and other

techniques tend to be scattered across disciplines, from theoretical computer science to signal processing to statistics.

Discovery and Data

Accelerating science, engineering, and manufacturing through AI methods requires large and diverse sources of data. At the same time, AI may hold the key to the limitations associated with that data. That is, applying data sources—from instruments, simulations, sensor networks, satellites, the scientific literature, and research results—is inherently challenging with respect to data being "FAIR" (findable, accessible, interoperable, reusable). AI systems can be employed to automate the creation of FAIR data and integrate it into knowledge repositories, in turn providing the architectural basis for new data infrastructure necessary to accelerate AI training and model development.

This high-volume data acquisition not only extends the end-to-end experimentation time but also limits experiments with time-sensitive phenomena. Smart data reduction techniques (e.g., filtering relevant data or point-of-interest data acquisition) will be necessities rather than features with the upcoming instruments such as those mentioned earlier.

Data produced by instruments, manufacturing systems, or engineered products (e.g., vehicles), often cannot be shared due to regulations (e.g., medical records or energy usage data) or the competitive nature of the data (e.g., factory or mobility data). AI-based federated learning techniques can accelerate model development, for instance, by harnessing proprietary manufacturing data from multiple sources. These techniques enable the development and training of models with data from many sources without requiring data sharing among them.

AI-based data services that leverage success to date of new DL and unsupervised learning techniques are vital to designing and operating increasingly large scale, complex infrastructure. These services will, in turn, require AI-based functions that can integrate and augment multimodal data sources including metadata, such as scientific instrument responses (e.g., flux and focus) in combination with a record of instrument configurations (e.g., motor positions, neutron chopper phases, monochromator bending parameters), and measurable instrument and environmental parameters (e.g., ring current, cooling water flow, and temperature). The integrated data will underpin AI services for developing generative models and decision-making functions that will be required to build advanced predictive models of accelerators, end stations, and sample delivery systems. Such services and models will also aid in automated alignment and calibration of instruments, stabilizing user operations, predicting and preventing catastrophic failures, and/or reducing the total downtime of the instrument.

While the infrastructure and methods needed to enable AI methods to access, learn from, and add to the broad body of knowledge are nascent, there are promising examples, such as the use of reinforcement learning, unsupervised learning, and classification techniques to automate labeling and creation of metadata.

Conclusions

Realizing the scientific capabilities discussed throughout this report will require extensive co-design work for domain scientists, facility designers, AI experts, mathematicians, computer scientists, and software research teams. Across the 16 chapters are scientific requirements that suggest a suite of new AI-capability building blocks and services, from design to control, augmented simulations to generative models, decision making to inverse problems, and the ability to learn not only from multi-model data (e.g., text, graphics, images,

waveforms, structured, time series) but from the domain knowledge embodied in the scientific literature.

To achieve the grand challenge of developing self-improving and self-adaptive hardware-software systems and applications, the services, applications, and software infrastructure must be both grounded by mathematical and AI foundations research and also implemented, evaluated, and adjusted over the coming decade. While this report is a not a detailed implementation plan, we can see possibilities for how to accelerate the opportunities identified by the community. One potential path is to partner with industry along at least two roadmaps.

The first is an "Instrument-to-Edge" activity that charts the course toward common tools and services for instrument, experiment, and infrastructure design, evaluation, optimization and steering, and safer operation across the DOE enterprise.

The second entails continuing efforts to advance a leadership computing, data, and analysis infrastructure that fully exploits and optimally supports new, AI-enabled software, data lifecycle, workflow, and modeling services and toolkits.

DOE's programmatic approaches, such as co-design or SciDAC programs, are ideal for developing the new AI services—packaged and supported as reusable toolkits and building blocks—that are required for self-driving laboratories and for steering scientific instruments. AI-based components including design, decision-making and evaluation, control and optimization, or the creation of generative models from instrument data and simulations are necessary to move from "AI *has potential for…*" to "AI *is enabling…*".

International leadership in AI over the coming decade will hinge on an integrated set of programs across four interdependent areas—new applications, software infrastructure, foundations, and hardware tools and technologies, feeding into and informed concurrently by DOE's scientific instrument facilities and by DOE's leadership class computing infrastructure.

This page intentionally blank.

01. Chemistry, Materials, and Nanoscience

The ability to design and refine materials and chemical compounds has always been key to the rapid advancement of society's technology and infrastructure. Today's complex technologies require a broad spectrum of needs when developing and optimizing materials and chemicals with desired performance [1–3], such as mechanical, electronic, optical, and magnetic properties (e.g., smartphones use up to 75 different elements compared to the twentieth-century version that had only ~30). This new level of technological complexity, combined with the need to search undiscovered areas of the chemical and materials landscape without clear theories or synthesis directions, [4] requires new paradigms that utilize artificial intelligence (AI).

AI will become an integral part of a scientist's arsenal, alongside pen and paper, and experimental and computational tools. It will accelerate the next scientific discoveries and the design and development of revolutionary technologies benefiting society. AI will identify both promising materials and chemicals, and the reaction pathways to make them [5]. Scientists will use AI to generate scientific data in a rational way, formulating new physical models and theoretical insights that drive new paths for rational design of materials and chemicals, and exploring atomic design spaces currently unimaginable.

1. State of the Art

Our ability to discover new materials and chemical reactions is driven by intuition, design rules, models, and theories derived from scientific data generated by experiments and simulation. The number of materials and chemical compounds that can be derived is astronomical, so finding the desired ones can be like looking for a needle in a haystack. Currently, various machine learning (ML) approaches are used to help scientists explore complex information and data sets with the goal of gaining new insights that lead to scientific discoveries. Future discoveries of advanced materials could be greatly accelerated through ML. Note, for example, the timeline from discovery of $LiMn_2O_4$ to nickel-manganese-cobalt (NMC) materials for batteries. Using known data, we could use ML to accelerate discovery of new material classes for batteries from 14 years to less than 5 years (Figure 1.1).

Nowadays, experimental characterization tools routinely provide picometer/picosecond resolved images at an ever-increasing rate, and, when coupled with a modern camera, are capable of providing several hundreds of frames per second. This pushes the data size into the several hundreds of terabytes (TB) per experiment for a single microscope [6]. Real-time analysis of this data, aided by AI, is

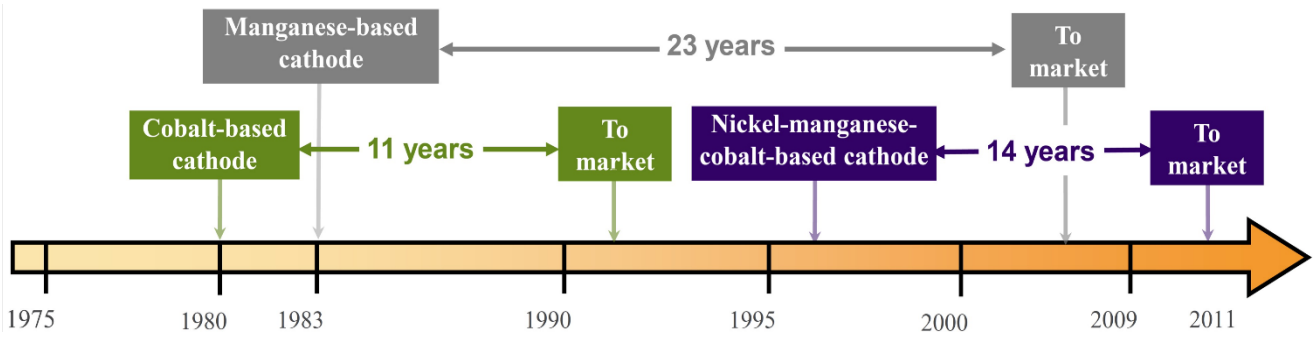


Figure 1.1 *Timeline from discovery of $LiMn_2O_4$ to NMC materials for batteries.*

needed to provide rapid feedback to and from models and simulations that can both inform and validate decisions. Such rapid feedback would also enable experimental adjustments on the fly. Progress has begun to address **two major gaps** in the current paradigm of materials design and discovery that typically proceeds via synthesis ⇒ characterization ⇒ theory.

First, continuous growth in high-performance computing (HPC) capabilities, combined with the development of efficient and scalable electronic structure calculation methods, is enabling scientists to virtually explore materials and chemical compounds. Large databases have come online containing the simulated properties of millions of relatively simple materials and chemical compounds. Deep learning (DL) approaches are being developed for various tasks, such as predicting properties or structure, but this barely scratches the surface of the full atomic design space available to us. Even more, the real world is far more complicated than the simple structures often studied by electronic structure calculations, and simulations investigating systems under device-relevant conditions are still prohibitively expensive. Advances are needed in reliable and precise computational techniques that accurately (and rapidly) address the increasingly complex functionalities required for today's technological applications.

Second, significant progress has been made toward fully exploiting all of the information contained in experimental and computational data to predict and understand new materials. An example is the automated image analysis and recognition based on DL networks that was successfully developed to identify and enumerate defects, and that created a library of (meta) stable defect configurations (Figure 1.2). The electronic properties of the sample surface were further explored by atomically resolved scanning tunneling microscopy (STM). Density functional theory (DFT) was used to estimate the STM signatures of the classified defects from the created library, allowing for the identification of several defect types across multiple imaging platforms. This approach now allows automatic creation of defect libraries in solids, explores the metastable configurations that are always present in real materials, and provides correlative studies with other atomically resolved techniques than can provide comprehensive insight into defect functionalities.

It is this integration and analysis of multiple, complex data sources combined with current state-of-the-art ML approaches that holds great promise for a drastic acceleration of materials and chemical compound discovery.

2. Major (Grand) Challenges

Finding new materials or chemical compounds that have unique properties needed for real-world applications—for example, batteries that hold **10x** the storage capacity compared to today's batteries, or materials that capture more solar energy at greater efficiency—is a grand challenge due to the nearly infinite chemical or atomic design space to which scientists have access. To date, our modern chemical and materials synthesis and discovery process incorporates a wide range of design rules and theories, alongside advanced characterization tools capable of observing synthesis processes on size and time scales at which they occur. At the same time, high-throughput screening via theory-driven approaches, per the materials genome, has provided guidance in identifying promising candidates optimized for particular properties. Early work in ML shows the potential for AI to start to provide guidance on the synthesis pathways to make a material or chemical. The underlying grand challenge as outlined by the Basic Energy Sciences Advisory Committee (BESAC) is how to design and perfect atom- and energy-efficient synthesis of revolutionary new forms of matter with tailored properties. This requires us to explore materials and chemical compounds compositions entirely unknown, driving questions such as, where in our atomic design space do we look? How do

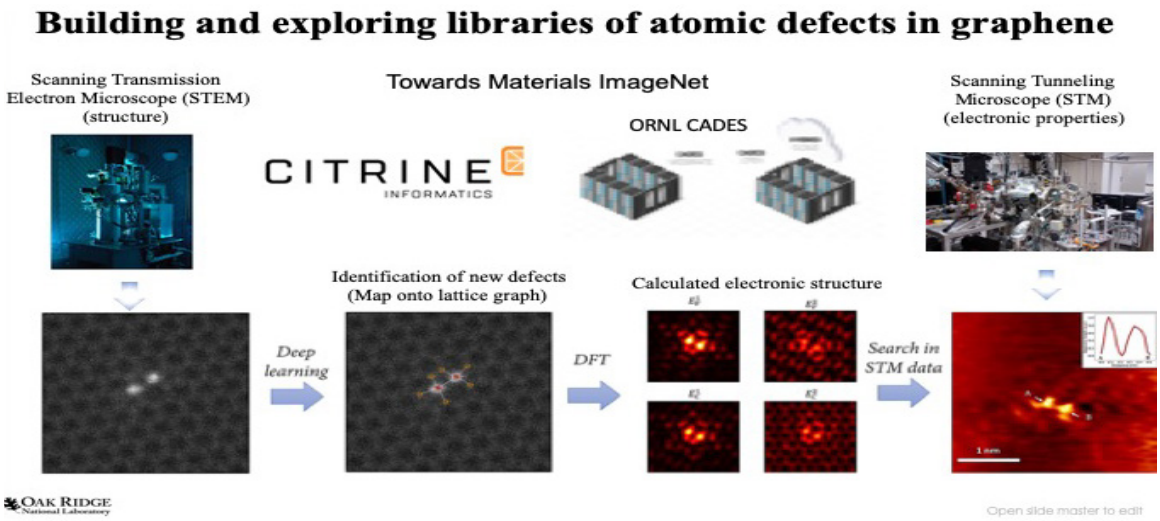


Figure 1.2 *A scanning transmission electron microscope (STEM) images materials where there are defects present or intentionally induced by the electron beam in the STEM. DL via convolutional neural networks is used to process the data to recognize and categorize defects. These data are populated into a database hosted by CITRINE Informatics. DFT calculations via HPC are used to predict STM images for the different defect classes, which then are used to train the DL in a similar fashion to the STEM, and then deposited into the database [7].*

we search the chosen space in the most efficient way or decide to move on to other areas? Can we develop new design rules? Aiding this would be the ability to understand the length- and time-scale evolution of functional chemical and materials systems.

The primary challenges are concisely described by BESAC's 2015 report, *Challenges at the Frontiers of Matter and Energy: Transformative Opportunities for Discovery Science*.

- Mastering Hierarchical Architectures and Beyond-Equilibrium Matter
- Beyond Ideal Materials and Systems: Understanding the Critical Roles of Heterogeneity, Interfaces, and Disorder
- Revolutionary Advances in Models, Mathematics, Algorithms, Data, and Computing
- Harnessing Coherence in Light and Matter
- Exploiting Transformative Advances in Imaging Capabilities across Multiple Scales

Specifically, gaps/challenges that need to be addressed by AI/ML are listed below.

Design metastable phases and materials that persist out of equilibrium. These materials enable access to a diversity of properties beyond the limits drawn by equilibrium thermodynamics. For example, optically driven processes of materials could provide more control over the chemical processes and lead to new materials, such as metastable phases or new low-dimensional materials with dynamics controlled by in-plane heterogeneity rather than layer stacking order. Another example is self-assembly, where transient (non-equilibrium) intermediate states frequently appear, and control of assembly pathways can enable improved structural control. Modern characterization systems such as electron and scanning probe microscopies may allow "bottom-up" fabrication of new structures that are metastable, which allows arrays, for example, of topological defects to be created with nanometer precision for desired properties. The challenge is to do this in an efficient and reproducible fashion; this requires in-line analytics and feedback of very high velocity and volume data streams.

In January 2018, the U.S. Department of Energy's (DOE's) Office of Advanced Scientific Computing Research (ASCR) hosted a Basic Research Needs workshop focused on ML for science. This workshop resulted in development of priority research directions (PRDs) for interpretability, domain awareness, robustness, and needed capabilities (Workshop report on Basic Research Needs for Scientific Machine Learning: Core Technologies for Artificial Intelligence, https://www.osti.gov/servlets/purl/1478744). Although the workshop highlighted significant investment in ML for the analysis of big data, there has been less activity on the generation of such data sets—a critical need as DOE's major experimental facility upgrades begin commissioning. PRD-6 from the workshop, intelligent automation and decision-support, is highly relevant as timely advances in AI and ML will be critical to enable the full scientific potential. To make AI/ML successful for the large experimental and computational data from our facilities, there are challenges in terms of archiving metadata and preserving provenance, workflows to manage data transfer to and from instruments and integration with HPC facilities, development of software stacks (federated), and uncertainty quantification to identify regions of model validity.

Understand and control interfacial processes and properties. Controlling interfaces (liquid/liquid, gas/solid, etc.) often rely on precise control of atomic bonding and molecular interactions between two dissimilar phases. The ideal strategy to avoid performance-limiting defects in materials, for example, is to minimize perturbation of the atomic order at the interface by preserving a high degree of crystallographic order (e.g., epitaxy). However, atomic scale insights into grown structures present significant inverse problems that have been difficult to address. This may potentially be tackled using combined physics-ML methodologies (Figure 1.3). Additionally, chemical separations, an area which is fundamentally important to almost every aspect of our daily lives, from the energy we utilize to our medications to chemical purification, including water, can see transformative advances with AI in terms of refining and optimizing experimental approaches. The use of AI will aid the pursuit of grand challenges such as understanding complex hierarchical correlations, from molecular-scale interactions up to transport phenomena, and mapping energy landscapes for the chemical and materials transformations that occur during aging of separation materials/chemicals.

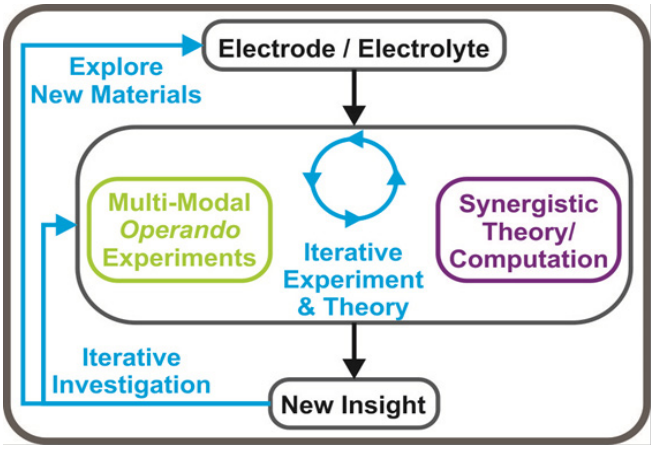


Figure 1.3 *An integrated approach for future design of materials interfaces tailored for performance. Key to this vision is inclusion of multi-modal operando experiments enabled by AI/ML.*

Design materials and molecules for quantum information sciences (QIS). Much of the transformative success of technologies underlying the information age was built on our ability to manipulate chemical composition and doping, and hence electronic band structure and electrochemical potential, within materials at tiny length scales, encode local electronic properties as the physical instantiation of information, and thus control the storage, flow, and processing of information. We now stand on the brink of a *quantum information revolution*. Here, breakthroughs will be driven by the ability to harness the interplay and evolution of quantum entangled and coherent ensembles as the physical representation and processing of information. This will provide radically new opportunities in computation, enabling exponentially higher speeds and efficiencies and the ability to solve problems that are currently intractable. As such, there is a desperate need to deliver systems for potential solid-state qubits, photon sources, and quantum sensing systems [BES Roundtable, Opportunities for Basic Research

for Next-Generation Quantum Systems, Oct. 30–31 (2017); Roundtable, Opportunities for Quantum Computing in Chemical and Materials Sciences, Oct. 31–Nov. 1 (2017)]. Promising advances at DOE facilities in layered materials stamping and a new pulsed laser deposition (PLD) system will generate rich structural, heterointerface, and functional property datasets that will require deep AI/ML analysis and real time control. This analysis/control will need to be done *in situ* and on the timeframe of the experiments to enable smart-steering of the synthesis processes toward successful quantum materials.

Understand the critical roles of heterogeneity in complex systems. Heterogeneities and interfaces underlie novel functionalities and drive dynamical processes, such as charge and exciton transport (e.g., along grain boundaries), charge separation (at Type II heterojunctions) and recombination (at Type I heterojunctions), spin evolution, and transport of ions or molecules through ordered and disordered systems (e.g., at battery interfaces or through metal organic frameworks). However, understanding transient and time-dependent processes in material and chemical systems is enormously challenging; examples include identifying chemical reaction pathways, visualizing electronic and optoelectronic processes at their native lengths (single atoms to many nanometers) and time scales (femto to nanoseconds and beyond) in heterogeneous materials, and studying exchange processes between excitations on various length scales. Progress can be made via high-throughput materials synthesis and automated atomic-scale/multimodal characterization. Here the aim is to broadly understand how population diversity influences growth and behavior, with the ultimate goal of creating a closed-loop materials property prediction, synthesis, and characterization loop. By understanding and controlling heterogeneity, it may be finally possible to design multifunctional and self-regenerating catalytic systems.

Understand and master energy and information with capabilities rivaling those of biological systems. Biological systems naturally transform and distribute energy through photosynthesis and subsequent decomposition of photosynthetic material. Conversion of energy to biomass can occur via various mechanisms, including photosynthetic and chemical pathways with oxygen (i.e., aerobic) and without oxygen (i.e., anaerobic). Greater insights are needed into the regulation of these pathways, the mechanisms responsible for the reactions, and environmental influences on the reactions. This improved understanding is a precursor to enabling changes in pathways that may uncover new or more efficient energy sources.

3. Advances in the Next Decade

In the next five to 10 years, AI will be an integral part of a scientist's discovery and design arsenal. Scientists will use AI to generate scientific data in a rational way, formulating new physical models and theoretical insights that drive new paths of rational design of materials and chemicals, exploring atomic design spaces currently unimaginable.

The ultimate form of AI for materials, chemistry, and nanoscience constitutes ***autonomous-smart experiments and simulations, including synthesis and automated discovery,*** that integrate all aspects of the materials and chemistry discovery loop—from preparation through characterization, to data interpretation and feedback—in order to minimize the experimental trials needed to achieve a desired property or set of properties. This could allow vastly more challenging materials and chemical compound problems to be tackled. However, such an autonomous process will still require *expert scientists in the loop* to ensure viability and success. Overall, the vision of "autonomous-smart experiments" is an as-yet unrealized grand challenge, as the parameter space is simply too large to manage in traditional ways. AI/ML can clearly be a

transformative key to bridge this gap, but it will require addressing a number of challenges (ranging from teaching the AI physical concepts and rational design decisions), making experimental instruments "smart," integrating experimental and simulation data, working with large and diverse sets of streaming data, and having precise control over the experiments. AI/ML can be transformative in terms of high-throughput screening, drastically accelerating simulation capabilities to achieve desired precision with very low computational cost and opening the door to virtually explore a much larger part of the available design space.

Efficient materials, chemical, and device characterization are critical elements in the scientific discovery workflow. As such, the characterization capabilities are constantly used for the determination of chemical composition, structure, physical properties, and overall functionality. In general, this involves (1) an analytical step to confirm that the target chemicals and/or materials are produced; (2) characterization of the physical properties, morphologies, defects, and interfaces of the functional materials and chemicals by multiple probes/techniques; (3) characterization of the functional properties, *in situ/operando*, in devices. This means it will require new analysis across all of these platforms, including registration of data from different instruments (e.g., pan sharpening) and scaling for structure-property mapping. It will be important to fully enable *in situ multimodal analysis with streaming data,* for example, implementing online analysis and active learning during an experiment when more than one type of probe is being used (as data will be streaming at potentially very high velocity and volume).

With AI and ML automation of model-building and decision-making in experimental loops, machine-guided synthesis, processing, and ultimately materials and chemistry discovery can be achieved, enabling discovery, synthesis, and control of novel processes and properties (Figure 1.4).

In the next decade, all the upgrades to DOE's light sources will be completed alongside the proton power upgrade at the neutron source. Thus, there will be significant advances and new information in the following areas.

New data sets/instruments online. There will be a continued increase in the capabilities in detectors/cameras alongside accelerators that will lead to a tremendous increase in potentially high-quality information from microscopes and light sources. Those instrument advances will provide extreme volumes and velocities of data that contain deep information regarding materials/chemistry processes alongside a modality that enables manipulation and control of the materials.

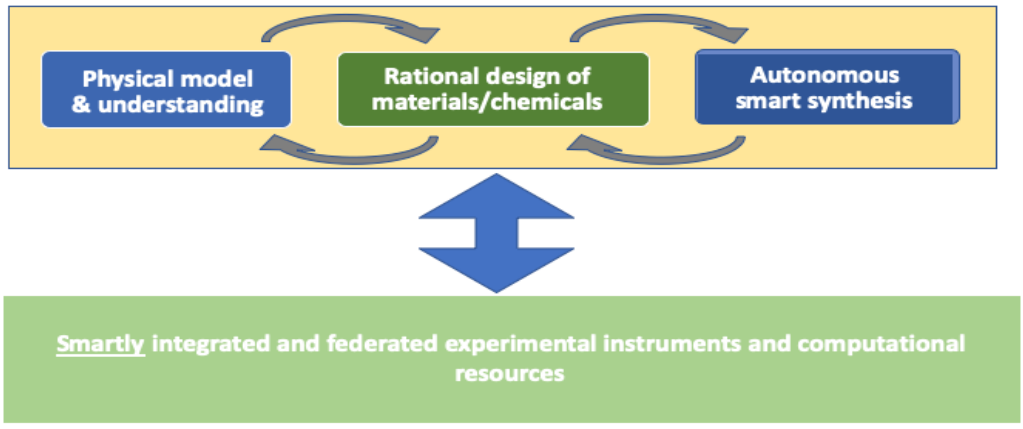


Figure 1.4 *Schematic illustration of the elements of experiments and computations that are required to enable autonomous-smart experiments for materials/chemical design/synthesis.*

Enhancement in big data and data curation. There must be a focused effort to link major facilities and capabilities, such as our leadership computing facilities and our microscopy, light, and neutron sources, to characterize and fully understand the new materials. We need a radical improvement on data sharing, analysis, and curation that will catalyze scientific discovery. *This requires the development of protocols, common data formats, and complete metadata to document and curate the full history and knowledge of the synthesized material.* Furthermore, workflows to integrate knowledge across multiple facilities, and the ability to create and draw on knowledge graphs to better inform modeling and propose new experiments, should be expected. Ultimately, a shared and curated source of data that is easily searchable and minable will be a fundamentally needed infrastructure. Progress is expected along the lines of new AI platforms that integrate diverse scientific data resources, including the literature, and respective mining engines, which will enable automatic development of training sets from heterogeneous experimental and simulated data (see Chapter 12, Data Lifecycle and Infrastructure).

Rare events detection and identification. Rare events are events that occur very infrequently, i.e., their frequency ranges from 0.1 percent to less than 10 percent. While these events are low probability, they can have high impact.

Events such as failure in materials under stress, or side reactions in gas phase chemistry that may occur on time scales too short for humans to observe, are very important to identify. Near-term adaptive control of some experiments—when implemented as real-time decision-making during an experiment—can identify regions of interest and save the relevant data. The introduction of AI into instrument control systems will allow detection when their alignment has drifted and then perform automated alignment and recalibration.

Computers and algorithms. There will continue to be major advances in computer capacity and mathematical algorithms which will further enhance the ability to perform in-line and real-time analysis of experimental and computational data.

Accelerated simulation. Continued advances in computing capacity and computational chemistry and materials methodologies, combined with ML network development, will provide new sets of data for AI/ML and decision making.

New AI/ML techniques. Advances are expected in reinforcement (algorithms that employ reward/punishment), active learning and neuromorphic computing that may be used “at the edge”—where people and things meet (AI/ML at the edge)—as well as in explainable and interpretable AI/ML (see Chapter 10, AI Foundations and Open Problems). Particularly important will be advances in AI/ML approaches that can deal effectively with sparse, unlabeled data.

4. Accelerating Development

To achieve the vision of autonomous-smart experiments/discovery, a number of technical challenges must be addressed. It will be critical to accelerate development in the following areas.

Advance edge computing and integrated experimental instruments. Computing at the experimental instrument(s) for on-the-fly analysis with feedback during an experiment will need to be implemented to maximize information gain and efficient control. This will be particularly important for multimodal experimental probes that require analysis across different platforms. Edge computing for automating aspects of experiments, such as for AI/ML-assisted tuning of the environment, importance sampling, next-experiment recommendation, etc., will be critical. Additionally, on-demand pipelines to HPC for automatic spawning of jobs directly related to

discoveries at the instrument are needed. This can be important for forming databases based on higher levels of ML models trained on simulated data, where the simulations would require an HPC environment. The goal is to provide fast *on-the-fly* analysis of "streaming" experimental data.

Enable *in situ* multimodal analysis. Characterization capabilities are constantly used for the determination of chemical composition, materials structure, physical properties, and how such properties correlate with functionality. In general, this involves (1) an analytical step to confirm that the target chemicals and/or materials are produced; (2) characterization of the physical properties, morphologies, defects, and interfaces of the functional materials, by multiple probes/techniques; (3) characterization of the multi-functional properties, *in situ*/*operando*, in devices (*in vacuo, in solute, in atmosphere*) across a broad frequency range. Achieving acceleration will require new *in situ* multimodal diagnostic approaches which incorporate all of these analytical platforms in one experiment. These include registration of data from different instruments/*in situ* probes and scaling (e.g., pan sharpening) for structure-property mapping, multimodal cross-correlation, and building of frameworks to integrate knowledge in a rigorous physics-based framework that incorporates uncertainty quantification meta data analytics. *In situ* data analytics, including cross modeling, will be approached on two levels, the first level at the point of experiment using edge computing and at the second level of HPC. The ML algorithms will be incorporated as a part of *in situ* multimodal analysis. It will lead to machine-guided decision-making algorithms for selection of optimal experimental condition, minimal number of experiments, and reduced model error.

Enable automated smart characterization. The use of active learning and Bayesian methodologies in combination with predictive modeling during experimental characterization can enable the efficient exploration of heterogeneities in materials and the delicate balance in chemical compounds and reactions. The goal is to minimize the uncertainty and to maximize physics knowledge gain.

Enable AI/ML approaches to represent physics. Dictated by the laws of physics, only discretized structures exist in nature. This "discreteness" needs to be represented properly in the encoding space to control erroneous predictions and misclassifications. New and novel mathematical approaches are needed to incorporate physical constraints and symmetries into the representation and encoding of chemical and materials data, feature detection, and the learning process itself. New kernels that can operate on hierarchical structured data for similarity quantification to enable the application of uncertainty-aware regression methods are also needed.

Enable big-fast data at the signal-noise edge. Use of ML models for characterization at the dose limited range is critical for autonomous experiments. Big-data-based techniques, such as four-dimensional scanning transmission electron microscopy (4D-STEM), are limited by how fast the data can be collected, with the bottlenecks arising from detector readout times and data transfer rates. This imposes constraints on the sample since it rules out beam-sensitive samples that will not be stable under the comparatively slower imaging conditions, and also dynamic *in situ* experiments. Fast detection is possible, but the data is noisier. Current state-of-the-art iterative analysis protocols are more susceptible to noise, and next-generation ML models trained on HPC-simulated datasets can be a way to bridge this gap between big data in microscopy and dynamic microscopy.

This includes integrating data efficiently from different characterization techniques to provide a more complete perspective on materials structure and function. Even with this promising progress, there is still tremendous need for work that can bridge a number of critical gaps,

including delivering a set of open-source petascale quantum simulation, data assimilation, and data analysis tools for functional materials design, within an approach that includes uncertainty quantification and experimental validation and verification of AI models (see Chapter 10, AI Foundations and Open Problems).

Develop a workforce that can work across domains. Existing and emerging training programs in chemistry and materials need to be expanded to ensure a workforce that understands AI approaches and how they can best benefit problems in chemistry and materials discovery.

5. Expected Outcomes

Success in achieving autonomous-smart experiments will lead to transformative advances in:

- The diversity of materials properties possible beyond the limits drawn by equilibrium thermodynamics or our imagination based on discovered design rules.
- The realization of multifunctional and self-regenerating catalytic systems.
- The control of interfaces optimized to perform desired functions.
- On-the-fly materials and (bio)chemical design and synthesis.
- The discovery of unknown synthesizable materials and complex chemical species **1000x faster** and with desired properties.

6. References

1. Riordan, M. & Hoddeson, L., Crystal Fire: The Invention of the Transistor and the Birth of the Information Age, W. W. Norton & Company, 1998.
2. Sze, S. M., *Physics of Semiconductor Devices,* 2nd Edition, John Wiley and Sons, New York, 1981.
3. Shockley, W., Electrons and Holes in Semiconductors: With Applications to Transistor Electronics, D. Van Nostrand Company, Inc., 1950.
4. Fuechsle, M. et al., A single-atom transistor. *Nat. Nanotechnol.* **7**, 242–246 (2012).
5. Sumpter, B. G., Vasudevan, R. K., Potok, T., Kalinin, S. V., A bridge for accelerating materials design*. npj Comp. Mat.* **1**: 15008 (2015). DOI: 10.1038/npjcompumats.2015.8
6. Kalinin, S. V., Sumpter, B. G., & Archibald, R. K., Big-deep-smart data in imaging for guiding materials design. *Nat. Mater.* **14**, 973–980 (2015).
7. M. Ziatdinov, et al., “Building and exploring libraries of atomic defects in graphene: Scanning transmission electron and scanning tunneling microscopy study,” Sci. Adv. 5:eaaw8989 (2019). DOI: 10.1126/sciadv.aaw8989.

This page intentionally blank.

02. Earth and Environmental Sciences

Earth and Environmental Sciences addresses some of the most pressing challenges in the nation, from natural resource utilization to maintaining our infrastructure and environment. In particular, recent events have highlighted the fact that our society is vulnerable to increasingly frequent natural hazards, including wildfire, drought and extreme precipitation events (Figure 2.1). An urgent need exists for improving our predictive capabilities of earth and environmental systems, including physical, chemical, and biological processes that govern the complex interactions among the land, atmosphere, subsurface, and ocean components from molecular to global scales, and from daily to decadal time scales.

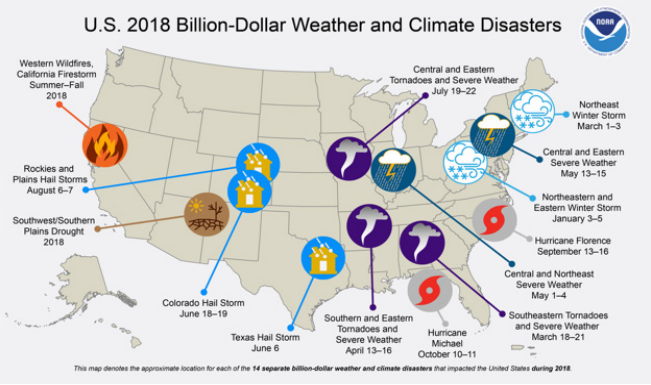


Figure 2.1 *Billion-dollar weather and climate disasters for the year 2018 [32].*

In recent decades, Earth observation capabilities have been revolutionized, based on a suite of novel sensor, analytics and telecommunication technologies. In particular, DOE has pioneered integrated observational capabilities at the laboratory scale (e.g., EMSL, SNS, ALS) and at field scales (e.g., NGEEs, ARM), as well as developed systems biology databases (e.g., KBase) and data archives (e.g., ESS-DIVE and ESGF). We now have access to several hundred petabytes of observational data of the Earth system in the U.S. alone; most of them in real-time. In parallel, predictive modeling capabilities have advanced significantly to simulate complex Earth systems, facilitated by HPC capabilities. Together, these vast observation and simulation data offer unique opportunities to apply AI approaches for improved understanding and scientific discovery in Earth and environmental sciences. AI methods offer the promise to accelerate development of advanced tools and the next generation of technology for assimilating observations and data-driven forecasting.

1. State of the Art

Applications of AI methods for Earth, environment, and climate research are in their infancy, but interest is growing rapidly as our ability to collect and create data outpaces our ability to assimilate, interpret, and understand it [24,3]. Primary applications include (1) knowledge discovery and estimation; (2) data assimilation and data-driven models; (3) model emulators, and (4) hybrid process-/ML-based models that integrate process scale data. Artificial neural networks (ANNs) and deep neural networks (DNNs) have been widely used for producing weather forecasts (e.g., [8]), spatiotemporal gap filling (e.g., [13]), and various remote sensing and geophysical image processing and analysis [3,21]. Random forest (RF) methods are widely used to understand and interpret complex environmental data [1], as well as to estimate environmental parameters such as soil properties at the global scale [12]. In addition, unsupervised learning and clustering methods have been used to discover key spatiotemporal patterns in large remote sensing and simulation datasets (e.g., [14]).

More recently, increasing interest in ML applications have fueled development of emulators for environmental process models, particularly in the subsurface and atmospheric sciences (e.g., [25,17,29]). New parameterizations based on ANNs have been developed for representing stochastic

convection based on simulations from fully resolved cloud models [6,23,22] with similar efforts for ocean modeling [4]. Earth, environment, and climate research is seeing rapid acceleration in the use of AI for data assimilation and for producing hybrid process-/ML-based models and physics-informed ML, including "active learning" methods and GANs [26,27].

2. Major (Grand) Challenges

Four grand challenges have emerged in the Earth, environment, and climate disciplines that could be revolutionized through application of AI methods and incorporation of burgeoning data, leading to new scientific discoveries and advances in energy security, national security, and adaptation and resilience to extremes in our changing environment.

Project environmental risk and develop resiliency in a changing environment. Increasing risks are posed by changing environmental conditions and increasing frequency of weather extremes on various aspects of our society and energy sector, including detrimental effects of wildfires, floods, droughts, wind, solar energy production and contamination (Figure 2.2). Our ability to assess the vulnerability from such changing conditions, mitigate imposed risks, and respond rapidly to such events is limited by the fidelity of modeling and observational tools. New advanced sensors coupled with edge computing capacity are now available for rapid data acquisition, but many challenges still exist for real-time data-model assimilation. New tools are needed to accelerate the projection of weather extremes and their result impacts on energy infrastructure and the built environment (i.e., buildings, roads, utilities) under changing environmental conditions. Efforts to build resiliency to address evolving risks will benefit from data-driven approaches that integrate smart sensing systems, built-for-purpose models, large ensemble forecasts to quantify uncertainty, and dynamic decision support systems for critical infrastructure. The 10-year goals include (1) development and understanding of predictive capabilities of Earth, environment, and climate models from sub-seasonal to decadal scales; (2) development of coupled datasets that are consistent across all components of the Earth, environment, and climate; (3) development of purpose-built and point-of-action forecast models of Earth, environment, and climate that are usable for

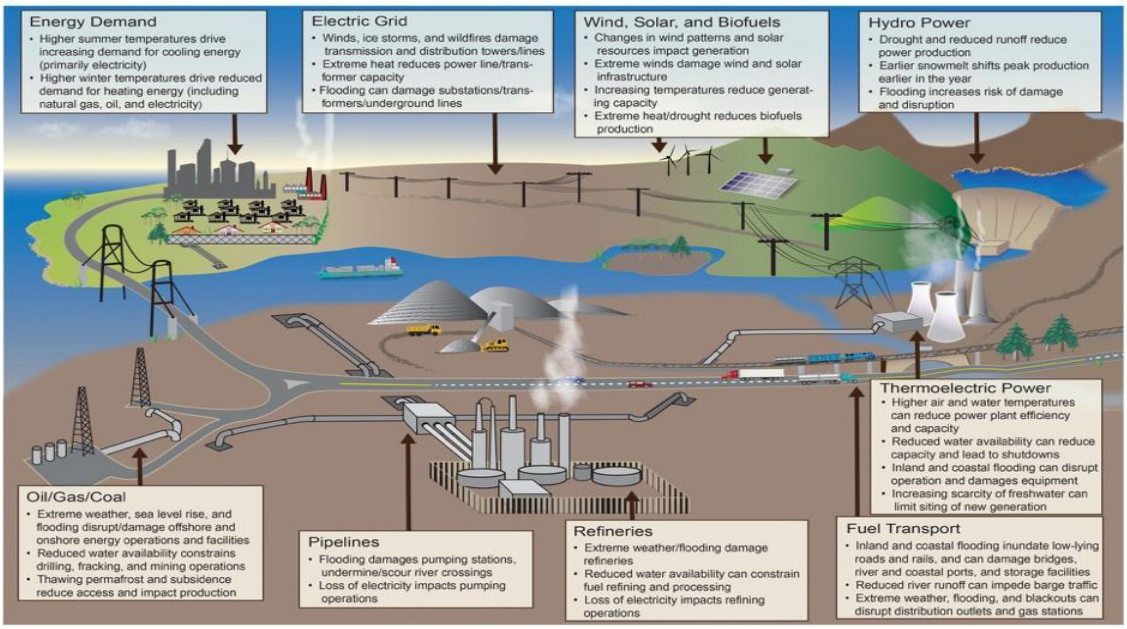


Figure 2.2 *There are many ways environmental conditions and changes in the environment affect energy systems [33].*

estimating risk and resilience, and (4) the scaling up of the observational capabilities of extreme events.

Develop adaptive subsurface management strategies for energy production and storage, and waste isolation. The energy security of our nation relies on the utilization of subsurface reservoirs for energy production and storage, carbon storage, and spent nuclear fuel storage. We need to substantially increase hydrocarbon extraction efficiency from unconventional reservoirs; discover and exploit hidden geothermal resources; reduce environmental impacts, including induced seismicity; dramatically increase geologic CO_2 storage; and improve prediction of the long-term fate and transport of contaminants. However, our capabilities to assimilate existing data to understand, reliably predict, and adaptively control subsurface processes are extremely limited (Figure 2.3). The subsurface datasets and real-time data streams are typically uncertain, disparate, diverse, sparse, and affected by scaling issues. The physical models of subsurface processes (e.g., flow, storage, stress, chemistry) are incomplete, uncertain, and frequently unreliable for making predictions. The 10-year goal would be seamless integration of multivariate data with real-time data streams into forecasts of system behavior with innovative visualization, including the capability for predictive models to test various hypothetical operational and economic scenarios, as needed, to guide operational decisions in near real-time.

Develop a predictive understanding of the Earth system under a changing environment. To advance the nation's energy and infrastructure security, a foundational scientific understanding of complex and dynamic biological, geochemical, and hydrological processes, and their interactions under environmental change, is required (Figure 2.4). The knowledge gained through this research must be incorporated into models of the Earth system—designed to simulate atmospheric,

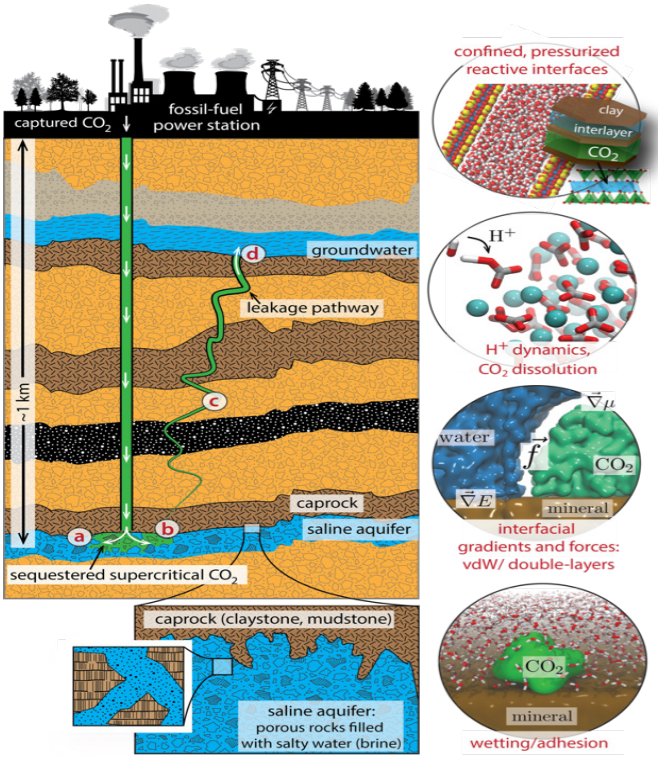


Figure 2.3 *AI/ML is required to connect multiscale data of geomechanical-chemical-transport trapping mechanisms in Geological Carbon Capture and Sequestration for the case of a deep saline reservoir [34].*

land surface, oceans, sea ice, land ice, and subsurface processes—to yield predictions of future climate and Earth system conditions under various scenarios of human factors such as population, socioeconomics, and energy production and use. Accurate predictions needed to close the carbon cycle require understanding the responses of terrestrial and marine ecosystems to changes in temperature and atmospheric composition and the feedbacks of those responses on the climate system. Integral to this research is characterizing the influence of water in mediating biological responses and in transferring energy, carbon, and nutrients across all components of the Earth system. In addition, leveraging advances in genomics and bioscience data promises to provide detailed understanding of plant/microbial functions and their adaptation and feedbacks to the changing environment. The 10-year goal would be leveraging AI methods for (1) assimilating large volumes of continuous observations into

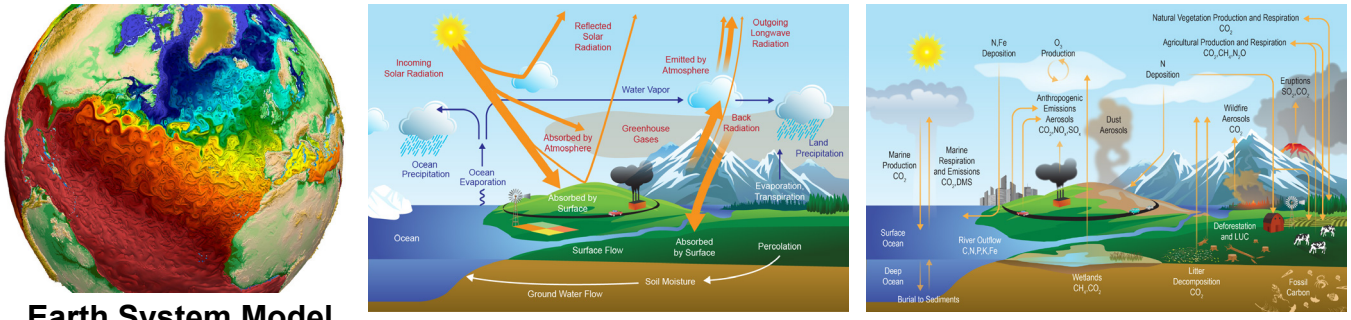


Figure 2.4 *Earth system models (ESMs) are designed to capture the behavior of interacting natural and anthropogenic processes and to project future behavior as a result of changes in population, economics and policy, and strategies for future energy production and use [31].*

data-driven models and for optimizing model parameters; (2) extrapolating sparse measurements across space and through time to characterize functional traits of biological systems and dynamic processes important for closing the carbon cycle, and (3) developing hybrid process-based/ML models that improve climate predictability and reduce uncertainty in future projections.

Ensure water security under a changing environment. Water resources are critical for human health, energy production, food security, and economic growth. The demand for fresh water is increasing because of the growing population and corresponding consumption practices. However, water availability and water quality are being impacted by climate change, extreme weather, and disturbances such as wildfire, droughts, floods, and land-use change. Processes affecting water quality and water availability span multiple spatiotemporal scales from soil microbiology to individual watersheds to continental scale hydrology (Figure 2.5). Therefore, water availability and water quality cannot be adequately addressed locally or regionally or within a single compartment. Methods are needed to integrate disparate and diverse multi-scale data with models of watersheds, rivers, and water utility infrastructure for near-real time prediction and water management. The 10-year goals

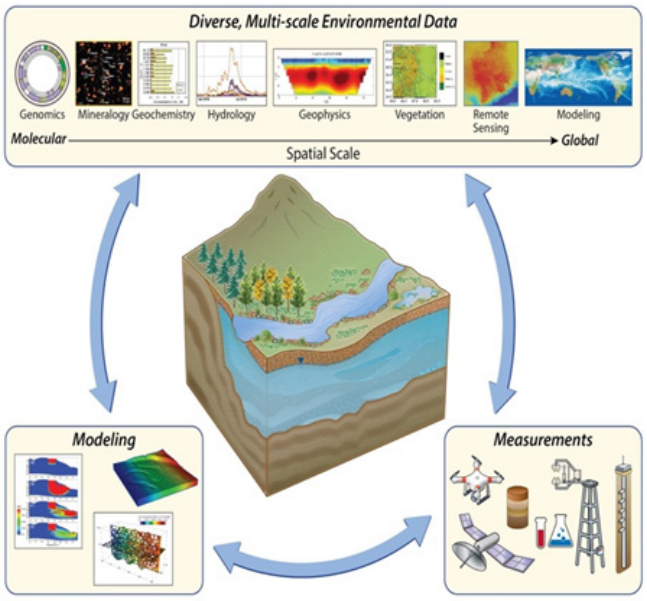


Figure 2.5 *Achieving predictive capabilities toward water security requires the integration of diverse data from multidisciplinary Earth sciences, including hydrology, ecology, climate, geology, geophysics, geochemistry, and microbiology. Adapted from [28].*

are (1) development and understanding of predictive capabilities of water availability and water quality at the continental scale; (2) development of modeling and sensing systems to obtain representative data across the range of scales and compartments; (3) development of scale-relevant theories to bridge scales for prediction and control, and (4) development of faster execution capabilities to predict water availability and water quality across scales.

3. Advances in the Next Decade

Observations of the atmosphere, biosphere, ocean, land, and subsurface are expected to improve considerably over the next five to 10 years. Increasingly, remote sensing data from satellite platforms is providing profiles through the depth of the atmosphere, oceans, and even soil layers over land and surface and subsurface deformation. This increases the volume of data available several fold. Ground-based measurements are also rapidly increasing in ubiquity and the variety of geophysical variables that are sensed. Geophysical methods continue to advance their capabilities particularly due to massive amounts of data collected by dense sensor deployments, or rapid advancements in fiber optic sensing. Observations collected on robotic platforms that navigate remote oceans and cheap sensors in everyday devices (Internet of Things) are becoming increasingly common for collecting environmental data (see Chapter 15, AI at the Edge). These types of sensors excel in the frequency of data collection and produce copious amounts of data. Data from the next generation of accelerators and light sources will be another source of large datasets in the next few years (see Chapter 14, AI for Imaging). Characterizing the properties of atmospheric particles, biogeochemistry of soil cores, and geochemical processes at nano and atomic scale in controlled laboratory settings will be widely available [20]. Recently, the ability to collect extensive, multimodal genomic and microbial data across spatiotemporal scales and tools enabling precise modification and control of biological (see Chapter 3, Biology and Life Sciences) and environmental systems have significantly expanded and are expected to enhance multifold in the future.

Below ground datasets are challenging to collect as most of the available data are based on *in situ* measurements, soil core collection, and subsequent laboratory analysis, and limited remote sensing technologies [16]. Synchrotron light sources, neutron scattering facilities, and electron beam imaging devices are becoming sufficiently penetrating that they allow us to observe directly the reactivity occurring inside a rock or other porous matrix [18,30]. Increasing flux has improved the time resolution dramatically; however, this creates a challenge in that the data sets derived are massive and unwieldy [11]. Artificial intelligence has the potential to allow us to quantify pore-fluid accessible surface areas of different mineral phases, fluid-fluid contact areas that allow us to measure residual trapping efficiencies for carbon dioxide. The increasing complexity of systems capable of being measured is directly matched by dramatic increases in the size and complexity of the associated data, necessitating new approaches to analysis and interpretation [9,7].

Simulation data sets are also increasing in size. Earth system modeling is a key application targeted by upcoming DOE exascale computing platforms. The resolution of Earth system models, such as the DOE E3SM model, is increasing toward resolving mesoscale extreme weather phenomena and eddy processes at the ice–ocean interface, and ultimately toward full cloud resolving models. Resolving detailed processes in these growingly complex models produces large increases in model output; however, I/O bandwidth limitations of high-performance computing platforms will increasingly limit these high-frequency, highly resolved data from being saved for later analysis. This points to both the need for online intelligent feature extraction to reduce simulation data [15], and to deploy ML and statistical analysis algorithms *in situ* within simulation codes to analyze output as it is being generated [2].

4. Accelerating Development

While the amount of data collected on the various components of the Earth system are increasing, the ability to use these data in developing improved forecasts and model

development has not kept pace. Numerical model development is constrained by the ability of scientists to evaluate the data, develop and test hypotheses, and produce new models. Earth and environmental data for global change presents a challenge to ML methods because the dimension of the data (e.g., spatial resolution) can be much greater than the number of data samples (e.g., time slices). Data are often multiscale, can be irregularly distributed (point cloud or unstructured mesh data), and in some cases can be sparse or missing in random ways (measurement bias). Data are usually correlated across large distances in space and time, which presents challenges for traditional ML methods that assume independent samples or that assume spatial regions can be analyzed independently. Computer science innovations will be required both in training algorithms and in distributed computation. Thus, to accelerate development, the following issues specific to environmental datasets will have to be addressed.

Multi-scale data. Earth and environmental data are often available from different sources (such as satellite sensors, *in situ* measurements and model simulations, and, increasingly, robotic sensors) and at varying spatial and temporal resolutions, exhibiting different characteristics (such as sampling frequency and accuracy).

Noisy, missing, and uncertain data. Earth and environmental data show different degrees of noise, incompleteness, and uncertainty. Satellite sensors can be noisy with clouds and snow cover; sensors may temporarily fail causing missing data; and some environmental variables can be measured only indirectly from other observations or model simulations with uncertainty.

Shortage of labeled data with ground truth. Collecting high-quality and high-resolution Earth and environmental data is very expensive and time-consuming, and for some environmental variables and processes (e.g., subsurface structure and subsurface flow) there are no ground truth observations.

Spatial and temporal heterogeneity. Earth and environmental processes have large spatiotemporal variability, which is highly correlated and structured. The data often have nonlinear relationships, feedbacks, non-stationary features, and low frequency high impact events.

Environmental forecasting is complex and uncertain. Environmental projections are developed using complex, coupled, nonlinear systems representing different components of the Earth system. This makes the projections from these models uncertain, with uncertainties propagated from data, model structural and model parameters. It is necessary to characterize these uncertainties and increase the credibility of the projections to support decision making.

What must we do to accelerate development?

a) Develop AI approaches to improve and optimize data acquisition, including sensor network optimization, data compression and edge computing (see Chapter 10, AI Foundations and Open Problems).

b) Establish the protocols and tools to allow access, transfer, curation, quality control, and maintenance of public datasets that can dynamically be coupled with the model/simulation systems (see Chapter 12, Data Life Cycle and Infrastructure).

c) Develop supervised, semi-supervised, and unsupervised AI systems for multiscale multi-type data (see Chapter 10, AI Foundations and Open Problems).

d) Relieve the bottlenecks in processing petabytes of data and speeding up the entire model development and training algorithms, by exploiting effective CPU/GPU communication patterns (see Chapter 13, Hardware Architectures).

e) Develop AI-enabled automated approaches for model development and hypothesis testing which can provide improved insights on physics, chemistry, and biogeochemistry (see Chapter 10, AI Foundations and Open Problems).

f) Develop fully AI/physics-coupled models that can ingest massive data and honor mass/energy conservations, and other physical principles (see Chapter 10, AI Foundations and Open Problems).

What are the top priorities?

- Develop AI-assisted data acquisition strategies associated with new robotics and *in situ* sensors.
- Develop consistent and high throughput data access, compression and transfer software for the variety of Earth and environmental science datasets.
- Apply automatic labeling and reduction of environmental datasets at various spatial and time scales.
- Advance explainability of AI approaches for modeling EC phenomena and avoiding the "black box" conundrum.
- Develop a hybrid approach to combine AI with physical principles for EC models, and develop robust explainable AI software for training and validating hybrid models (Figure 2.6).
- Develop robust and consistent protocols for testing the transferability and reproducibility of AI models across a wide range of conditions.
- Advance uncertainty quantification methodology as an integral part of the AI workflow.

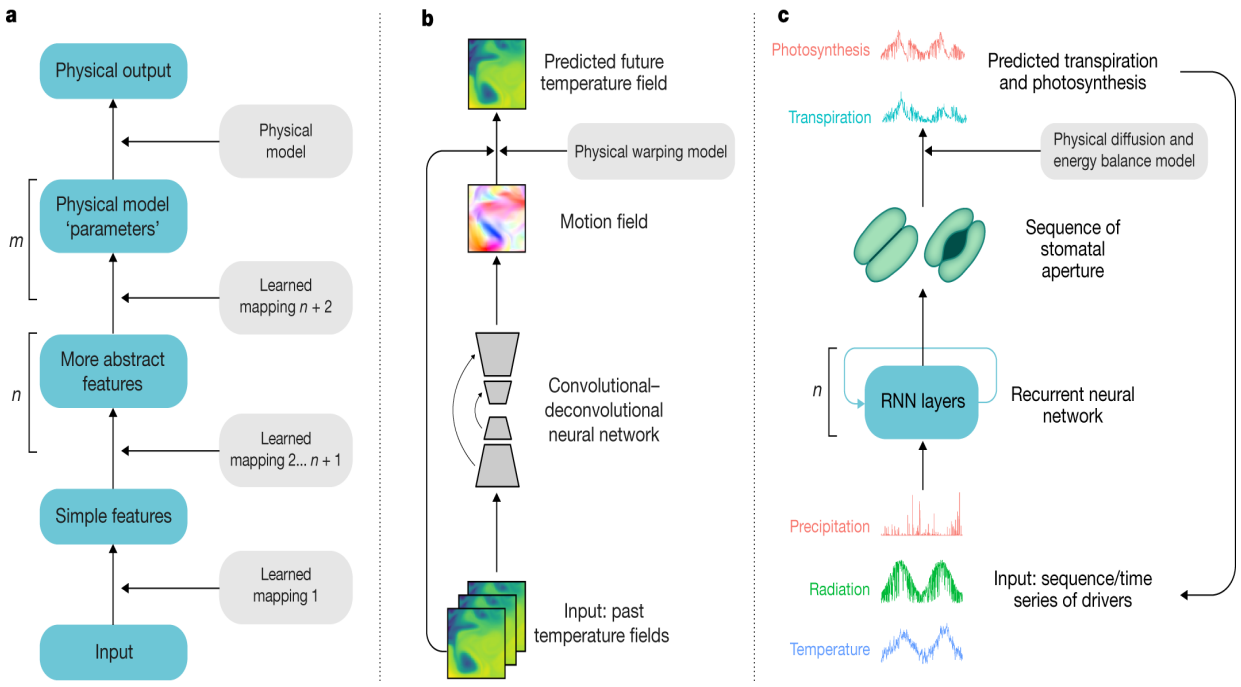


Figure 2.6 *Hybrid approach that combines AI with physical understanding to address some of the black box issues and make the models physically consistent: (a) shows a multilayer neural network, with* n *the number of neural layers and* m *the number of physical layers;* b *and* c *are concrete examples of hybrid modelling; (b) prediction of sea-surface temperatures from past temperature fields; (c) a biological regulation process (opening of the stomatal 'valves' controlling water vapor flux from the leaves) is modelled with a recurrent neural network [24]. Hybrid models are useful for replacing poorly understood or unresolved (sub-grid scale) phenomena. Challenges include (a) obey physical constraints; (b) quantify uncertainties in the parameters in the network models; and (c) develop methods for adding explanation to the network models and parameters. Training hybrid models using offline or online methods need exploration.*

How do we improve scale?
Most of the applications in the literature were performed with small datasets in the range of gigabytes. Handling large data throughputs will be necessary to fully realize the potential of AI, requiring scaling up of computational infrastructure and the ability of the AI algorithms to handle large volumes of data. As these data volumes grow over time, data cannot be kept in memory continuously for retraining ANNs. Thus, we need AI algorithms that scale in terms of intelligence while processing very large data volumes out-of-core.

Scalability will be an important challenge, potentially requiring a move toward streaming analysis methods adapted to spatially and temporally correlated data. When conducted online in conjunction with physics simulations, additional scalability challenges will arise due to incompatibilities between traditional AI distributed training techniques and distributed computation for physics simulations, requiring new, potentially domain-specific algorithms.

5. Expected Outcomes

Success in these areas means:

- AI will revolutionize the development of process scale models by accelerating the process of discovery and model creation.
- AI will enable rapid prototyping of purpose-built models of Earth system processes and energy/built infrastructure that will enhance national energy and water security preparedness.
- AI will make it feasible to merge large datasets with numerical models for a new generation of predictive models that can span the forecast scale from daily to decadal and local to global.

6. References

1. Basu, S., Kumbier, K., Brown, J. B., & Yu, B. (2018). Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences*, *115*(8), 1943-1948.
2. Baydin, A. G., Shao, L., Bhimji, W., Heinrich, L., Meadows, L., Liu, J., & Ma, M. (2019). Etalumis: Bringing Probabilistic Programming to Scientific Simulators at Scale. *arXiv preprint arXiv:1907.03382*.
3. Bergen, K. J., Johnson, P. A., Maarten, V., & Beroza, G. C. (2019). Machine learning for data-driven discovery in solid Earth geoscience. *Science*, *363*(6433), eaau0323.
4. Bolton, Thomas, and Laure Zanna. "Applications of deep learning to ocean data inference and subgrid parameterization." *Journal of Advances in Modeling Earth Systems* 11, no. 1 (2019): 376-399.
5. Brantley, S. L. (2018) Shale Network Database, Consortium for Universities for the Advancement of Hydrologic Sciences, Inc. (CUAHSI). DOI: 10.4211/his-data-shalenetwork
6. Brenowitz, N. D., & Bretherton, C. S. (2018). Prognostic validation of a neural network unified physics parameterization. *Geophysical Research Letters*, 45, 6289-6298. https://doi.org/10.1029/2018GL07851
7. Cherukara, M. J., Nashed, Y. S. G., & Harder, R. J. Real-time coherent diffraction inversion using deep generative networks (2018). Scientific reports **8**(1), 165230.
8. Collins, W. & P. Tissot. An artificial neural network model to predict thunderstorms within 400 km^2 South Texas domain, *Meteorological Applications* 22, no. 3 (2015): 650-665.
9. Deng, J. et al.. Correlative 3D x-ray fluorescence and ptychographic tomography of frozen-hydrated green algae (2018), *Sci. Adv.***4**(11) eaau4548(1-10).

10. Flinchum, B. A., et al. Critical Zone Structure Under a Granite Ridge Inferred From Drilling and Three-Dimensional Seismic Refraction Data. (2018) J. Geophys. Res.: Earth Surf. 123 (6), 1317-1343.

11. Godinho, J. R. A., Gehrke, K. M., Stack, A. G., & Lee, P. D. (2016) The dynamic nature of crystal growth in pores. *Sci. Rep.,* 6:33086. DOI: 10.1038/srep33086

12. Hengl, T., et al. (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PLoS one*, *12*(2), e0169748.

13. Krasnopolsky, V., Nadiga, S., Mehra, A., Bayler, E., & Behringer, D. (2016). Neural networks technique for filling gaps in satellite measurements: Application to ocean color observations. *Computational Intelligence and Neuroscience* (2016): 29.

14. Kumar, J., Mills, R. T., Hoffman, F. M., & Hargrove, W. W. (2011). Parallel k-means clustering for quantitative ecoregion delineation using large data sets. *Procedia Computer Science*, *4*, 1602-1611.

15. Kurth, T. et al. (2018) Exascale deep learning for climate analytics. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis*, pp. 51. IEEE Press,.

16. Laanait, N., He, Q., Borisevich, & A. Y. Reconstruction of 3-D Atomic Distortions from Electron Microscopy with Deep Learning. *arXiv, [cond-mat.mtrl-sci]*arXiv:1902.06876v1 19 Feb 2019

17. Liu, Y., Sun, W., & Durlofsky, L. J. (2019). A deep-learning-based geological parameterization for history matching complex models. *Mathematical Geosciences*, *51*(6), 725-766.

18. Li, Z., et al. (2016) Searching for anomalous methane in shallow groundwater near shale gas wells. *J. Contam. Hydrol.* 195, 23-30. DOI: 10.1016/j.jconhyd.2016.10.005

19. Lin, H.W., Tegmark, M., & Rolnick, D. Why Does Deep and Cheap Learning Work So Well? *J Stat Phys* (2017) 168: 1223.

20. Ling, F. T., et al. (2018) Nanospectroscopy Captures Nanoscale Compositional Zonation in Barite Solid Solutions. *Sci. Reports,*8:13041. DOI:10.1038/s41598-018-31335-3

21. Nogueira, K., Penatti, O. A., & dos Santos, J. A. (2017). Towards better exploiting convolutional neural networks for remote sensing scene classification. *Pattern Recognition*, *61*, 539-556.

22. O'Gorman, P. A., & Dwyer, J. G. (2018). Using machine learning to parameterize moist convection: Potential for modeling of climate, climate change, and extreme events. *Journal of Advances in Modeling Earth Systems*, 10, 2548–2563. https://doi.org/10.1029/2018MS001351

23. Rasp, Stephan, Michael S. Pritchard, and Pierre Gentine. "Deep learning to represent subgrid processes in climate models." *Proceedings of the National Academy of Sciences* 115, no. 39 (2018): 9684-9689.

24. Reichstein, M., et al. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195.

25. Scher, S. (2018). Toward Data-Driven Weather and Climate Forecasting: Approximating a Simple General Circulation Model With Deep Learning. *Geophysical Research Letters*, *45*(22), 12-616.

26. Schneider, T., Lan, S., Stuart, A., & Teixeira, J. (2017). Earth system modeling 2.0: A blueprint for models that learn from observations and targeted high-resolution simulations. *Geophysical Research Letters*, 44, 12,396-12,417. https://doi.org/10.1002/2017GL076101

27. Tartakovsky, M., C. Ortiz Marrero, P. Perdikaris, G. D. Tartakovsky, and D. Barajas-Solano, "Learning Parameters and Constitutive Relation- ships with Physics Informed Deep Neural Networks," arXiv e-prints, p. arXiv:1808.03398, Aug 2018.

28. Varadharajan et al., "Launching an Accessible Archive of Environmental Data," Eos, vol. 100. 2019.

29. Wang, J., Balaprakash, P., and Kotamarthi, R. (2019). Fast domain-aware neural network emulation of a planetary boundary layer parameterization in a numerical weather forecast model, Geosci. Model Dev., 12, 4261–4274, https://doi.org/10.5194/gmd-12-4261-2019.

30. Zachara, J., et al. (2016) Internal Domains of Natural Porous Media Revealed: Critical Locations for Transport, Storage, and Chemical Reaction. *Environ. Sci. Technol.* 50, 2811-2829 DOI: 10.1021/acs.est.5b05015

31. Hoffman, F. M., et al. (2017. International Land Model Benchmarking (ILAMB) 2016 Workshop Report, Technical Report DOE/SC-0186, U.S. Department of Energy, Office of Science, Germantown, Maryland, USA, doi:10.2172/1330803.

32. https://www.ncdc.noaa.gov/billions/

33. http://www.energy.gov/downloads/usenergy-sector-vulnerabilities-climate-change-andextreme-weather

34. Zarzycki, P. Towards understanding of Reactive Interfaces in Geological CO2 Sequestration, RIGECO, ERC-2015-CoG Proposal 682274, September 2015.

03. Biology and Life Sciences

The capacity to predict, control, and understand biological systems in mechanistic, often molecular detail, is on the horizon. Biology is being transformed by the ability to collect large, multimodal data across spatiotemporal scales, as well as by tools that enable precise modification and control of biological and environmental systems.

Concomitant advances in data analysis, ML, and new hardware architectures, coupled with HPC-enabled simulations are transforming our capacity to connect molecular interactions to higher levels of organization, from cells to ecosystems. To deliver on the promise of emerging technologies—to offer personalized medical solutions by developing and testing mechanistic hypotheses in tractable laboratory or *in silico* settings—requires fundamental advances in statistical ML and AI that integrate massively multiscale and multimodal sensing modalities.

1. State of the Art

The dawn of AI-enabled discovery in biology has occurred. Population genomics data is being used to learn the bases of complex traits, enabling researchers to discover non-linear molecular and gene-regulatory interactions along with the architecture of their phenotypic manifestations [1]. Elsewhere, neuroscientists are learning the dynamics of the thousands of neurons that control behavior from electrical and imaging data [2]. Synthetic biologists are building workflows that automate the inverse design of microbial and plant cells [3]. Computational biologists are using AI to learn force fields to enable near-exact molecular dynamic (MD) simulations with fully quantized electrons and nuclei [4]. Such analyses were intractable only a few years ago, and now the pace of innovation driven by AI technologies is accelerating.

However, realizing the future potential of AI-enabled bioscience is impeded by limitations of the computational learning frameworks that exist today. AI must be predictive of complex phenomena and simultaneously provide insight into the underlying biophysical processes they model [5]. Analyses that enable understanding have something in common: they are amenable to human exploration, statistical inference, and model discovery and selection. For example, a perfect, atomistic generative model of a bioreactor could be available, yet, if that model is not amenable to goal-based optimization—to inverse design—its utility is limited to “guess and check” prediction. In biology, guess-and-check prediction is often useless (if we had a strong hypothesis about the system as a whole, we wouldn’t be resorting to AI in the first place). Furthermore, if the model is as complex as the system itself, have we really learned how it works, or merely demonstrated the ability to replicate it *in silico*?

These challenges are particularly clear in healthcare, one of the fastest growing segments of the digital universe, which is expected to reach 2,314 exabytes of data by 2020 [6]. While the average lifespan in the U.S. (79 years) has increased 30 years over the past century, medical research has been less successful at prolonging *healthy* life (i.e., health span). Prolonging our lifespan without prolonging our health span is financially unsustainable for our nation (total costs of age-related diseases are expected to skyrocket, exceeding $1.5 trillion in the U.S. by 2030). AI could offer powerful solutions to these challenges by enabling powerful utilization of rapidly accumulating health data. This ambitious endeavor requires data-driven mapping of the human genome (i.e., genomic profile), phenome (i.e., physiologic status), and exposome (i.e., physical and social environment) in real-time and across the human lifetime. It is clear that the state of the

art falls far short of what the economy and society requires to survive.

2. Major (Grand) Challenges

Biological systems are dynamic processes characterized by combinatorically vast configuration spaces and the presence of emergent control principles at multiple levels of spatiotemporal organization. The overarching challenge before us is to enable the mechanistic characterization of biological systems through increasingly automated cycles of multimodal observation followed by experimentation. We see this manifest in three grand challenges.

Build the capacity to design custom biological systems capable of addressing major global health and environmental challenges – "build life to spec." Synthetic biology leverages engineering approaches to produce biological systems to a given specification (e.g., producing a target drug [7] or the capacity to invade cancer cells [8]). Tools are available that promise to disrupt this field: clustered regularly interspaced short palindromic repeats (CRISPR)-enabled genetic editing, high-throughput multi-omics phenotyping, and exponentially growing DNA synthesis capabilities, among others. However, synthetic biology will only reach its full potential when we have developed the capability to predict the behavior of biological systems, to develop first principles models, and to observe biological systems with much finer spatial and temporal resolution [9]. AI can provide the required predictive power, and improved methods for interrogating fitted AI learners may ultimately facilitate the detailed mechanistic understanding needed to support synthetic biology (Figure 3.1).

"Digital twins" of organisms will be a key enabling technology toward the capacity to design organisms to specifications—quantitatively modeling and simulating the behavior of complex biosystems [10] (see also

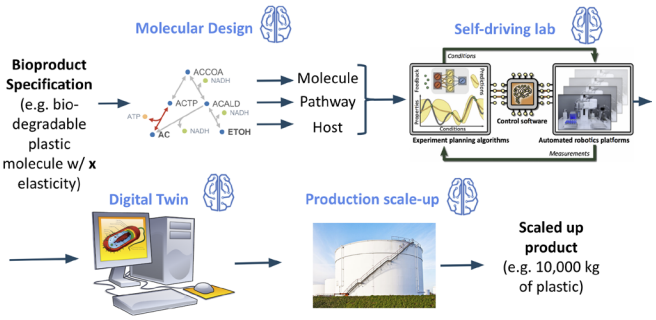


Figure 3.1 *AI can revolutionize synthetic biology if applied wisely. AI can help systematically choose molecules that fit a desired specification, and propose possible pathways and hosts to synthesize it. AI can help power self-driving labs able to collect high-quality, abundant data needed for ML to be effective. AI can complement mechanistic models to accurately simulate and model cells in a variety of environments. This can make production scale-up a predictable endeavor, a process that is presently more an art than a science.*

Chapter 7: Engineering and Manufacturing, and Chapter 11: Software Environments and Software Research). Digital twins of cells could be created by combining traditional mechanistic models with AI algorithms, leveraging the predictive capabilities of the latter and the insight of the former. Importantly, the ability to modify cells through synthetic biology tools brings about the possibility of validating and sequentially constraining such models systematically. It is reasonable to anticipate that obtaining first-principles models for biological systems will soon be on the horizon.

Foundational challenges remain—even the "vocabulary" of biological systems is resplendent with "dark matter" [11]. Despite progress in the past decade, the fundamental challenge of systematically exploring small-molecule chemical space to find new applications and biological knowledge remains largely unsolved. At least a third of sequenced genes across organisms are of completely unknown function. In meta-metabolomics experiments, it is rare that more than 5 percent of mass-spectra can be identified. To understand how organisms operate in ecological and environmental contexts, learning the molecular vocabulary of life is a prerequisite. Using AI to integrate multi-omics

data constitutes an opportunity to accelerate the discovery of function for these "dark" molecules.

Bioimaging technologies are rapidly improving in resolution and dynamic range. CryoEM tomography enables atomistic modeling of complex macromolecules [12]. However, extant tools for learning these models from low signal-to-noise cryoEM data rely heavily on real-time inputs from human operators. Work stations for state-of-the-art microscopes must be proximal to the scopes themselves for guidance during experiments derived from human insight and intuition. Beyond cryoEM, hyperspectral imaging provides increasing power to discover biomarkers and elucidate chemistry without appeal to destructive omics modalities. With advances in AI, real-time imaging of biochemical processes and landscapes in living samples is on the horizon [13]. Machine vision for low signal-to-noise technologies, as well as tensor-on-tensor regression strategies to translate between hyperspectral and omics modalities are needed to radically increase the automation, throughput, and discovery-power of bioimaging technologies.

The modeling of biomechanical systems pose similar challenges—and opportunities. For example, vascular flow simulations to understand the fluid dynamics that result in aneurysms and other anatomical anomalies are poised to deliver early prognosis of patient risks that, today, are rarely detected prior to pathogenesis. Coupling physical simulations to AI "hypervisors" to guide variable mesh resolution stands to radically accelerate the modeling of complex biophysical systems. The same technologies will enable the study of fluid dynamics in bioreactors, or cell-free systems for chemical or pharmaceutical production—where understanding fluid dynamics and diffusion is essential to achieving efficiency.

More generally, using AI to design organisms to a given specification requires large amounts of high-quality data. We cannot produce these data without leveraging automation. The codesign of algorithms and automated systems for data collection therefore arises as a need rather than a luxury [14] (see "self-driving labs" below).

Learn to systematically manage and engineer global environmental systems by obtaining a predictive understanding of ecosystems and their services. Attempting to understand how ecosystem services emerge from organismal and environmental interactions is central in environmental and biomedical sciences. The mechanisms behind carbon, nitrogen, phosphorous, potassium, and micronutrient cycles are determined by the integration of microbial, plant, fungal, metazoan, and viral interactions that, despite decades of quantitative ecology, remain challenging to predict, or even quantify. In recent years, our capacity to measure the ecology, chemistry, and hydrology that give rise to nutrient dynamics has evolved exponentially—metagenomics, untargeted chemistry, hyperspectral imaging, satellite imaging, *in situ* sensing, and soon quantum sensing systems. The challenge is discovering mechanistic models that are amenable to inverse design, thus enabling intervention at scales relevant to engineering our troposphere.

With growing global demand for fuel, food, water, and predictable weather, learning to engineer ecosystems has become urgent. In the U.S. alone, there are more than 1.1B acres of managed lands [15]. In the last 100 years, some 50% of soil carbon has been depleted through land use practices and soils [16]. Before us is the unprecedented opportunity to transform our managed lands into engines for environmental control. Atmospheric carbon is rising—an opportunity to mine this carbon presents a trillion dollar opportunity—to enrich our soils with labile carbon, enhancing the fertility, and therefore the value of our farmlands, to render marginal lands fertile, to grow our economy, and to feed our future population. In our depleted soils, prescriptions of chemical fertilizers are overused, which pollute our fresh and marine waters, leading to

algal blooms and marine dead zones. The soil-water interface is equally important, and currently extremely difficult to simulate or model with any accuracy—and it is essential to understand if we are to intelligently manage marine and freshwater algal blooms, and to ameliorate marine dead zones.

AI technologies can reveal the emergent controls of these enormously complex systems, and enable us to engineer our environment to radically expand the range of arable lands, while improving our freshwater availability and quality—in part, by replacing our dependence on chemical fertilizers with designed plant and microbial biosystems. The modeling of macroecology, and cognizance of impacts on species distributions and clines, are required for the intentional design and engineering of ecological processes. Such models may ultimately reveal control principles for natural ecosystems, enabling the responsible stewardship of our managed wildlands. Moreover, these ambitions require rigorous biosecurity, which itself is a design problem ripe for AI-powered guidance. AI for biosecurity will be under exquisite scrutiny—these applications are likely to push the development of secure, explainable AI systems with rigorous statistical guarantees.

Integration of AI with experimentally constrained, large-scale, biophysically detailed simulations will be required to refine and construct forward and inverse models—a particular challenge in the biological and environmental sciences, where most knowledge is stored only in the literature. Novel methods are required for extracting and organizing knowledge in constructs compatible with guiding learning in AI architectures, resulting in biologically meaningful discoveries.

Throughout the biological and environmental sciences, forward and inverse models of meso/macroscale measurements need to be constructed that provide multiscale understanding of cellular and community functions. These capabilities are required to predict, control, and understand the biological processes underlying productivity, health, disease, and bio-resilience to environmental conditions.

AI technologies aimed at ecosystem control have enormous implications for human health and biomanufacturing as well. For example, the challenge of efficiently scaling up reactor results from lab-scale (50 ml) to commercial volumes (10,000 l) requires understanding the biodynamics that lead to stable production. The need to identify and understand meaningful levels of organization in biological systems from a control perspective is exceptionally clear in the biomedical sciences. Diseases are caused by small-scale disruptions (e.g., genetic mutations) that manifest at larger scales. Effective medical treatments require identification, prediction, and control of biological processes. However, most small scale processes are currently immeasurable in humans at the required time-scales (Figure 3.2). For example, in the brain, while meso/macroscale measurements (e.g., electrocorticography (ECoG), functional magnetic resonance imaging (fMRI)) have revealed principles of global processing of brain areas in humans, the precise biophysical mechanisms that relate these signals to the activity of individual neurons is unclear. This impedes translation between basic neuroscience findings and our understanding of the human brain in health and disease, including dementia. To overcome these challenges, AI is needed that can discover nonlinear “governing equations” from high-dimensional, noisy time-series data with unobserved influences to bridge the gap between observed processes and those that require control.

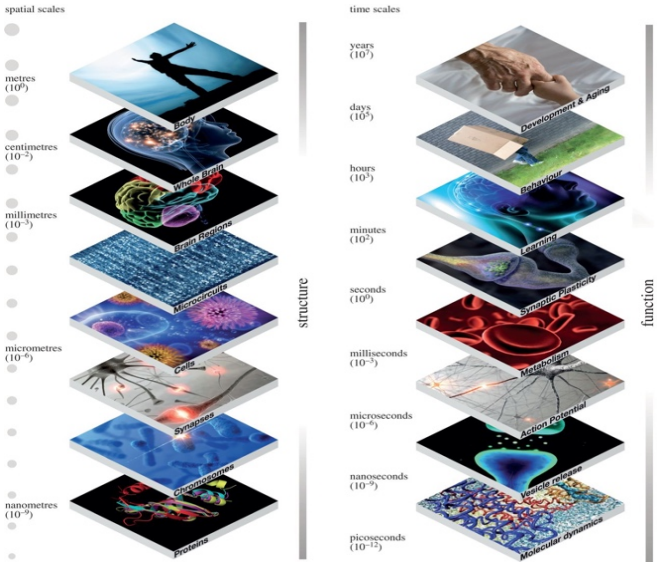


Figure 3.2 *Biological systems, including humans, constitute the integration of many levels of spatiotemporal organization. AI technologies hold promise to enable the systematic discovery of the manifestations of molecular interactions and processes on higher levels of physiological organization.*

Develop AI-enabled, self-driving laboratories to enable game-changing advances in the understanding and deployment of biological, chemical, and environmental systems. Fundamental to the role of AI in Science, and in particular biological, chemical, and environmental sciences, is the advancement of laboratories through automation and decision support (Figure 3.3).

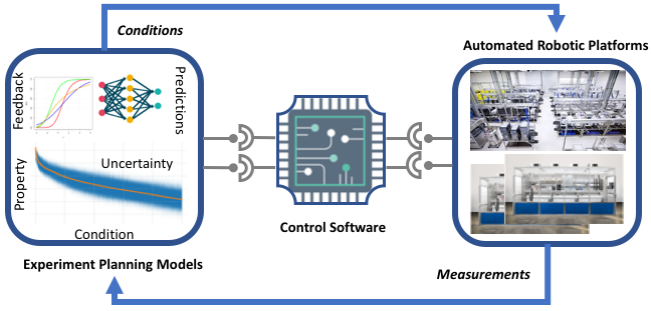


Figure 3.3 *AI-enabled self-driving laboratories couple automated robotics platforms for experimentation and data collection, with AI systems that choose not only the parameters for the next experiment but also the hypotheses to be tested. Figure adapted from Häse et al., Trends in Chemistry 2019.*

However, laboratory automation without carefully guided experimental design will contribute to the aggregation of low-value data. As in the manufacturing sector, we will soon see the effects of automation and robotics throughout biology. Industrial robots featuring high-quality computing capabilities, improved operational mobility, and machine vision systems are needed for future laboratories, particularly in synthetic biology, where goals will include genetic engineering toward an optimized design specification.

The future of these highly automated laboratories coupled with autonomous robots with enhanced dexterity must be intricately connected with the advancement of our application of AI to the challenges outlined above. Self-driving laboratories will require tight coupling with advanced AI models capable of representing complex biology far beyond what is possible today. Current AI approaches, such as model validation, uncertainty quantification, and active learning are relatively immature and will need to be common throughout our science to drive the execution of laboratory experiments—for example in molecular biology reactions, chemical reactions, and high resolution imaging—to a continuous feedback loop of data in the coming years.

3. Advances in the Next Decade

In the next 10 years, it will be possible to automate the process of biological discovery on unprecedented scales. The promise of self-driving laboratories is exceptional, and may underlie our capacity to achieve many other grand challenges: AI algorithms that design optimal experiments to reduce model uncertainty and constrain their own constructs toward learning mechanisms and robotically perform the experiments, the improvement of reducibility and reductions in cost, and time-to-discovery. Equally importantly, AI amenable to inverse design will enable “hypothesis discovery” and reduce our collective reliance on human intuition, with the potential to accelerate the pace of biological and environmental sciences by orders of magnitude.

However, there are two significant advances that need to be achieved. Because much of modern biology is not in the “big data” regime, model training must become more data-efficient than it is today. Additionally, scientifically meaningful insights from the fitted AI must be extracted.

First there is the need for data efficiency. Alternative learning approaches need to be developed that do not require highly overparameterized models for optimization, but that still admit the capabilities of neural networks—the ability to extract hierarchical representation from raw data (essential when data inputs lack semantic meaning, as with imaging data) and exquisite generalization accuracy. Here, biology may provide inspiration for AI methods development: biomimetic systems are needed that radically expand the domain of transfer learning and one-shot or few-shot learning. For the foreseeable future, there will remain scientific regimes in which dozens or hundreds of observations derive only from enormous community efforts. In ecosystems biology, for example, metagenomics and other molecular surveys will likely remain ultra-sparse at landscape scales. In biomedical sciences, in which phase 1 clinical trials are an essential data point in the lifecycle of a novel therapeutic or procedure, methods amenable to “small data” regimes will continue to be required.

Second, scientists will need methods to extract scientifically meaningful insights from what an AI model has learned from the data. Two complementary approaches are emerging. In one approach, human-understandable reduced order surrogate models (ROSMs) are extracted from more complicated models that accurately represent what an AI algorithm has learned during training. In a second approach, scientific knowledge and constraints are imparted to the architecture or objective function to ‘focus’ the learned representations so they are scientifically interpretable. There has been initial success in both the physical and biological sciences in this direction. The mathematical foundations of constrained representation learning as it relates to the geometry of loss-surfaces during training (hence the learnability) and inference (hence the generalizability) need substantial attention. While these initial steps are promising, much more work is required in these and other areas of AI.

To obtain complete descriptions of what class-leading deep neural networks learn from data during training will require new mathematics. This is a daunting prospect for a community built on yoking tools from statistics, numerical optimization, linear algebra, and dozens of other areas, not on developing novel theory. Here, biology may be as useful to the future of AI as AI is to biology. There is an opportunity to draw inspiration from the remarkable adaptability and self-regularization of biology to produce the next generation of AI algorithms and hardware. Blue sky research into alternative learning automata that reduce the initial ambient dimension of high-performing learning architectures is urgently needed for applications in the biological sciences. New learning “atoms” will undoubtedly come with new hardware requirements, and blue sky research has the potential to advance in both directions simultaneously.

4. Accelerating Development

Biological datasets must scale in their quantity, quality, and provenance. We need increased standardization of measurement techniques and metadata collection across the biosciences, and reconceptualized data sources as streams rather than the result of single experiments. The lack of data is by far the largest threat to the dawn of strongly AI-enabled biology.

Further, data availability faces special challenges in the biomedical sciences. We must establish the infrastructure required to make communal use of data that cannot be moved or revealed due to privacy concerns. An outstanding issue for sensitive domains, such

as health and medicine, is how to preserve privacy while computing with shared data to obtain insights. Removing personal identifiers and confidential details is insufficient, as an attacker can still make inferences to recover aspects of the missing data. Inference attacks can also jeopardize AI algorithms over shared data by targeting the shared AI model training process and the trained model itself. Indeed, serious threats are encountered in collective AI endeavors that aggregate data from different sources, since the most vulnerable source establishes the overall security level. This is an underdeveloped field of AI research in which Research & Development (R&D) investments are well warranted to develop new solutions so that the community can responsibly and privately share sensitive data for aggregated analysis, including training shared AI models, and performing transfer learning with sensitive data.

With the support of other federal agencies, the DOE national laboratories could provide a secure environment for objective benchmarking of AI algorithms against community consensus metrics to detect, monitor, and possibly correct dataset biases or inconsistent AI technology performance. First, investment is needed in foundational technologies to promote a rigorous statistical framework to monitor for potential biases or inaccuracies in collected data. During the deployment phase, rigorous quality control should be implemented, monitoring AI performance across subgroups to confirm robust performance or identify performance gaps.

5. Expected Outcomes

Throughout the biosciences, ultimately, the expected outcome is an understanding of life, from the ground up.

- Mining excess carbon from the atmosphere, revolutionizing human health, and engineering microbes and ecosystems to given specifications is within reach. Success in the development of AI for biology can transform our farmlands into an engine for soil security and the economic development of rural America. AI has the potential to extend the average human life, while significantly reducing healthcare costs.
- The potential impacts of AI technologies for health are difficult to overstate. Studies estimate that every federal dollar invested to map the human genome returned $60–$140 to the U.S. economy [17]. By leveraging federal health data assets, DOE's computing capabilities, and AI, novel solutions can be developed to extend health span and rein in costs by understanding the broad spectrum of factors impacting well-being and discovering cost-effective approaches to scale promising precision medicine solutions.
- Impacts on synthetic and environmental biology will become increasingly apparent as the AI technologies are developed to understand how ecosystem services emerge from biological processes. AI capabilities, coupled with retrobiosynthesis tools from synthetic biology pinpointing genetic and molecular controls of complex traits, can dramatically change the time scales for product realization, whether that product is a biofuel, a soil amendment, or the foundational understanding of a natural ecosystem.

6. References

1. Garcia, B. J. et al. Phytobiome and Transcriptional Adaptation of *Populus deltoides* to Acute Progressive Drought and Cyclic Drought. *Phytobiomes Journal*. (2018) **2**(4), 249-60.
2. Bouchard K. E., et al. Union of Intersections (UoI) for Interpretable Data Driven Discovery and Prediction. *Advances in Neural Information Processing System*. (2017) **30**:1078-86.

3. Lawson C. E., et al. Common principles and best practices for engineering microbiomes. *Nat Rev Microbiol.* 2019. Epub 2019/09/25. doi: 10.1038/s41579-019-0255-9. PubMed PMID: 31548653.
4. Chmiela, S., et al. Towards exact molecular dynamics simulations with machine-learned force fields. *Nat. Commun.* **9**. 3887 (2018).
5. Murdoch, W. J. et al., Interpretable machine learning: definitions, methods, and applications. arXiv preprint. 2019.
6. Harnessing the Power of Data in Health. Stanford Medicine Health Trends Report. 2017.
7. Paddon, C. J., et al., High-level semi-synthetic production of the potent antimalarial artemisinin. *Nature* **496** 528-532 (25 April 2013).
8. Anderson, J. C., et al. Environmentally controlled invasion of cancer cells by engineered bacteria. *J. Mol. Biol.* 355(4):619-27 (27 Jan 2006).
9. Gardner, T. S. Synthetic biology: from hype to impact. *Trends Biotechnol*. **31**(3):123-5 (2013 Mar).
10. Gambhir, S. S., et al. Toward achieving precision health. *Sci. Transl. Med.* **10**(430) (28 Feb 2018).
11. Blaser, M. J., et al. Toward a predictive understanding of Earth's microbiomes to address 21st century challenges. *Am. Soc. Microbiol.* (2016) doi: 10.1128/mBio.00714-16.
12. Allegretti, M., et al. Horizontal membrane-intrinsic α-helices in the stator a-subunit of an F-type ATP synthase. *Nature* **521**, 237-240 (14 May 2015).
13. Hermes, M., et al. Mid-IR hyperspectral imaging for label-free histopathology and cytology. J. Optics **20**(2) (24 Jan 2018).
14. Carbonell, P., T. Radivojevic and H. G. Martin. Opportunities at the Intersection of Synthetic Biology, Machine Learning, and Automation. *ACS Synth. Biol.* 2019, 8, 7, 1474-1477 (19 July 2019).
15. Census of Agriculture: Summary and State Data. United States Department of Agriculture. 2007.
16. Lal, R. Soil carbon sequestration to mitigate climate change. *Geoderma* **123**(1-2):1-22 (Nov 2004).
17. Hood, L. and L. Rowen. The Human Genome Project: big science transforms biology and medicine. *Genome Med*. **5**(9):79 (13 Sep 2013).

04. High Energy Physics

High Energy Physics (HEP) is concerned with discovering the ultimate constituents of matter and uncovering the nature of space and time. The underlying theory and associated experiments cover the smallest scales in all of science to the very largest. In the DOE context, this research quest is divided into three Frontiers: Cosmic, Energy, and Intensity [1].

The Cosmic Frontier uses probes relying on multi-wavelength surveys of the sky. The probes treat the universe itself as an experimental apparatus to investigate the mysteries of dark energy and dark matter—the primordial fluctuations from which all cosmic structure came to be—and to determine the masses of neutrinos, the lightest known material particles. In addition, experiments searching for direct evidence for dark matter fall within the purview of the Cosmic Frontier.

The Energy Frontier studies the fundamental constituents of matter by accelerating and colliding charged particles at very high energies in particle accelerators and by recreating conditions that only existed in the very early universe. The massive detectors that are used to study the collision events are among the most complex scientific devices ever constructed by humans (Figure 4.1). Work in the Energy Frontier is centered on searches for physics beyond the particle physics Standard Model, and investigation of the properties of the Higgs boson, discovered in 2012.

Intensity Frontier experiments require very sensitive detectors to study rare processes, and intense particle beams are often needed for this purpose. The primary area of interest here is the neutrino sector. Neutrinos are known to exist in three types (‘flavors’) and change flavor via quantum oscillations as they propagate in space and time. The oscillations imply the existence of neutrino mass. The origins of neutrino mass, the mass ordering,

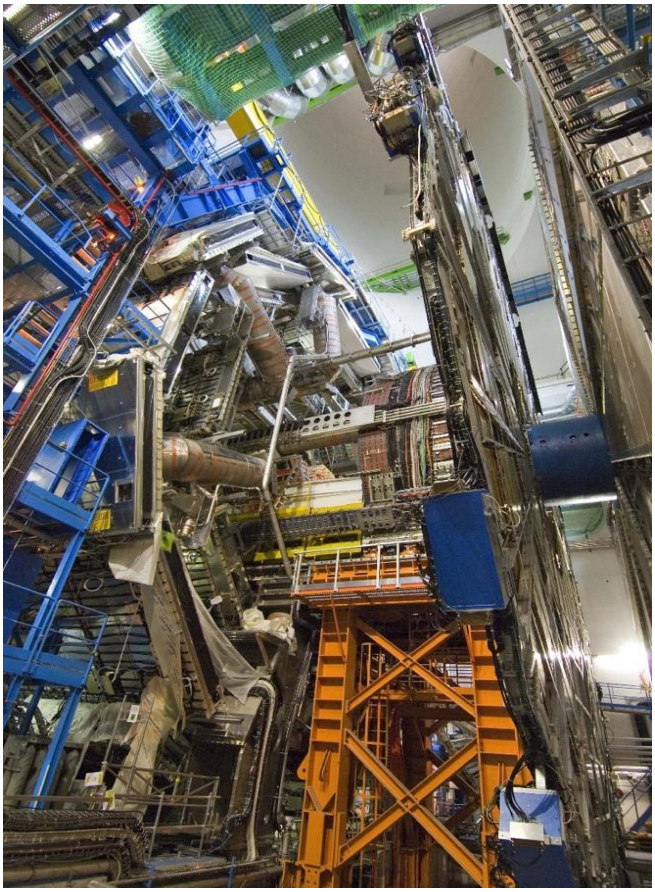


Figure 4.1 *The ATLAS detector at the LHC under construction in 2007.*

and whether neutrinos are their own anti-particles are just a few of the questions being addressed by Intensity Frontier experiments.

A defining characteristic of all experiments in this field is the generation of large, complex datasets that can range from the hundreds of petabytes to exabytes. In addition, simulation data, required to interpret the experiments, can reach similar scales. The experiments also feature high data throughputs. Because of both the volume and velocity of data, AI approaches are needed at multiple levels in the data management chain to improve the understanding of subtle systematics effects and to open new avenues for scientific discovery (see Chapter 10, AI Foundations and Open Problems).

The deployment of AI techniques in HEP has much in common with strategies and use cases discussed in other chapters. Certain general notions such as automated discovery, end-to-end workflows, explainability, and integration of data and theory are, of course, ubiquitous. More specifically, the idea of digital twins (see Chapter 8, Smart Energy Infrastructure) strongly resonates with the modeling-intensive approach characteristic of HEP. Data curation (see Chapter 12, Data Life Cycle and Infrastructure) is an essential aspect of HEP science. AI with edge systems (see Chapter 15, AI at the Edge) is analogous to HEP detector online computing tasks. Finally, employing AI in reconstruction and tracking is highly applicable, as HEP could profit from advances in physics-informed AI models for sparse, high-precision measurements (see Chapter 10, AI Foundations and Open Problems).

1. State of the Art

Advanced statistical methods and classical ML approaches have a long and productive history in particle physics, and crowd-sourcing techniques have been put to excellent use by cosmologists to lead to new discoveries. There are, therefore, a great number of natural applications of AI methods, a large fraction of which can potentially exploit the burgeoning activity in deep learning. Though in early stages, many ideas are being actively investigated with a view to addressing a number of crucial problems.

Cosmology offers multiple challenges being tackled today using AI approaches. Examples can be found in areas such as (a) photometric redshift estimation [2], (b) image analysis and feature extraction [3], (c) reconstruction methodologies [4] (including gap-filling), (d) object [5] and real-time transient classification, (e) inference frameworks [6], and (f) fast predictions derived from expensive simulations (emulators) [7].

The AI methodologies employed are as broad as the problems to be solved. They range from deep learning and active learning methods to random forest classifications, and they include more traditional machine learning approaches such as Gaussian process modeling. A noteworthy feature is the close connection with statistics—in particular, sampling theory and Bayesian methods— because of a focus on topics such as detailed verification and validation, which are typically not considered in non-scientific applications.

The Cosmic Frontier provides a rich application area for several reasons (Figure 4.2). First, cosmology is based on large observational datasets rather than on isolated experiments. The observational nature of the field makes it oftentimes impossible to extract the full information content from the data without the use of optimized learning algorithms. Image analysis approaches to disentangle images of galaxies (deblending), analysis of photometric data for redshift estimation, and feature extraction to identify (e.g., strong lenses) are just a handful of examples that have been actively developed in cosmology.

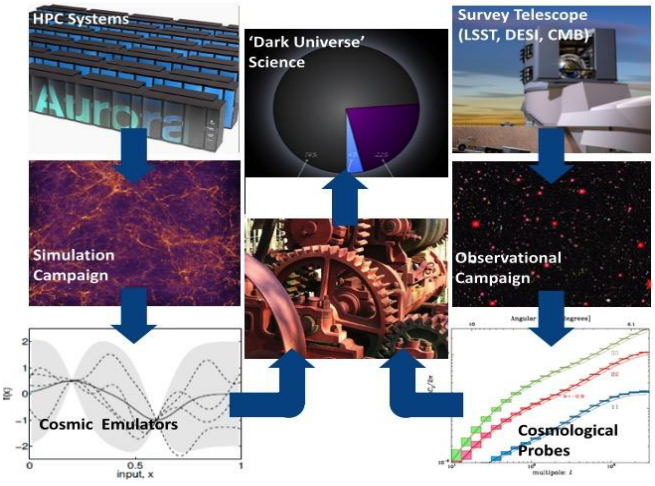


Figure 4.2 *Cosmological inference problem: AI methods will contribute in all individual phases as well as in a full end-to-end analysis.*

Second, object classification plays a critical role in cosmology and AI offers manifold approaches in this area. In particular, AI has

been used for the classification of transient objects, such as supernovae.

Third, the ultimate goal of cosmology is to infer the underlying physics of the universe from a complex set of data products across multiple wavebands spanning many orders of magnitude. This endeavor has to necessarily combine sophisticated data analysis with the best possible, and very computationally intensive, cosmological simulations. Here, ML approaches have been successfully used to develop precision emulators, and to help mitigate systematic effects. The already successful ongoing efforts using AI in cosmology along with new—and possibly unexpected—approaches will come together in the next 10 to 15 years to revolutionize our understanding of the universe and help answer some of the deepest questions in physics.

ML-based methods already define the state-of-the-art in a number of areas in particle physics experiments, including event and particle identification and energy estimation. The dominant algorithms are boosted decision trees and neural networks [8,9]. Since model training is the most computationally expensive component, particle physics experiments are increasingly employing sophisticated ML applications and reaping high value from them through the rapid turnaround of training and optimization tasks by spiking large scale, but relatively brief, workloads into a variety of large-scale computational resources, including the ASCR LCFs and NERSC, other HPC systems, Grids, and Clouds. Highly scalable distributed workload management systems already exist due to the use of the Grid paradigm in particle physics, and these systems (e.g., PanDA and HEPCloud) can be used to provide an integrated capability that runs seamlessly over a heterogeneous resource environment.

Accurate detector simulation using known interactions is necessary to compare with actual data in order to search for new physics. This complex task involves modeling the events via event generators followed by detailed, time-consuming Monte Carlo simulations for the interactions within the detectors. ML techniques for replacing the slow pieces of the simulations hold significant future promise, and work on these is currently underway. Event generators have a large number of parameters and tuning these in high-dimensional space is another obvious AI application (e.g., using Bayesian optimization [10]). ML techniques have been used for a long time to reconstruct certain characteristics of collision events from detector raw data [11] via pattern recognition and classification methods (including boosted decision trees and neural networks).

More recently, unsupervised or weakly supervised anomaly detection models (e.g., [12]) have been applied to model-independent resonance searches, opening new opportunities to detect physics Beyond the Standard Model (BSM). ML applications also have a place in theoretical approaches, such as the estimation of parton distribution functions, which cannot be computed from first principles QCD alone, and which need to be determined using experimental data [13].

AI techniques have been used successfully in the Intensity frontier experiments NovA [14] and MicroBooNE [15], which employed convolutional neural networks (CNNs), as these are particularly suited for applications to the large, homogeneous detectors that are characteristic of neutrino experiments. These techniques have been shown to outperform algorithms used previously, in part because they can exploit the suitability of GPUs for ameliorating the high training costs of CNNs.

2. Major (Grand) Challenges

The grand challenges in high energy physics, described as follows, are driven by the availability of high-volume, high-throughput data with significantly enhanced scientific value in resolution, sensitivity, and physics coverage.

Reconstruct the history of the universe using AI techniques. During the next decade, rich data sets will appear from advanced survey telescopes. At the same time, the advent of exascale computing will enable the next generation of sophisticated cosmological simulations, modeling structure formation in unprecedented detail. The new observations—unparalleled in depth and resolution at the observed scales—combined with the simulations and advances in AI, will allow the reconstruction of the history of the universe from the Big Bang until today, from the largest scales down to our own Galaxy. We will advance our understanding of the nature of dark energy and dark matter, gain insight to the earliest moments of the universe as currently described by inflation, and measure the mass of the neutrino. AI will play a pivotal role in this endeavor. Conventional methods, such as 2-point correlation function measurements, fail to extract all of the information encoded in the data. To optimally extract information while maintaining robustness, new AI techniques combined with statistical methods and HPC simulations will need to be developed. This combination will enable predictions deep into the nonlinear regime of structure formation, spanning a large mass and spatiotemporal dynamic range. Not only will this sharply determine the cosmological parameters casting light on fundamental physics, it will enable us to run the movie of our own universe back to the far past—the era of primordial fluctuations—as well as forward, enabling a glimpse into the future evolution of our local universe.

Advance knowledge of cosmic structure formation with the AI-driven Automated Cosmology Experiment (ACE). Based on advances in the next decade and driven by observational facilities such as the Large Synoptic Survey Telescope (LSST) and the Dark Energy Survey Instrument (DESI), the next generation of cosmological surveys will enable a new approach to cosmology via a fully automated, AI-driven, cosmological experiment, ACE. By combining already available cosmological data with unprecedented simulations generated by exascale and beyond computing capabilities, AI will enable a fully optimized experimental set-up. This set-up will include (1) the development of an optimal observing strategy, given the scientific focus of the observations (for example, finding the best compromise for deep versus wide field observations to deliver the highest-accuracy dark energy constraints); (2) best-possible methods to remove systematics from the data; (3) increased processing speed; (4) optimal calibration, and (5) the search of new observable features and anomalies and, therefore, the identification of new observing opportunities. ACE would be a combination of survey telescopes with follow-up instruments to enable fast tracking of unexpected objects as well as transient follow-ups important for cosmology. The continuous analysis of the data stream in combination with the predictions from the simulations will allow further on-the-fly optimization of the survey. ACE will be the cosmological equivalent of the Event Horizon Telescope (EHT) [16], a concerted effort between a network of radio telescopes that captured the image of a black hole and its shadow for the first time. In a similar way, ACE will capture the structures in the universe in a concerted effort of optical telescopes to shed light on the dark universe.

Zettascale AI to uncover new fundamental physics. Over the next decade, we will deploy AI-controlled, city-size scientific instruments (particle accelerators and particle detectors) that produce zettabytes of detector data. AI-powered hardware will filter the detector data in microseconds. AI-simulations of the detector response will enable high-precision studies, while completely unsupervised AI-searches for "New Physics" will open new windows for discovery (Figure 4.3).

To make this vision reality, we need:

i) Intelligent Operations. AI algorithms for anomaly detection will monitor the performance of particle accelerators, detectors, and computing systems, looking for early signs of

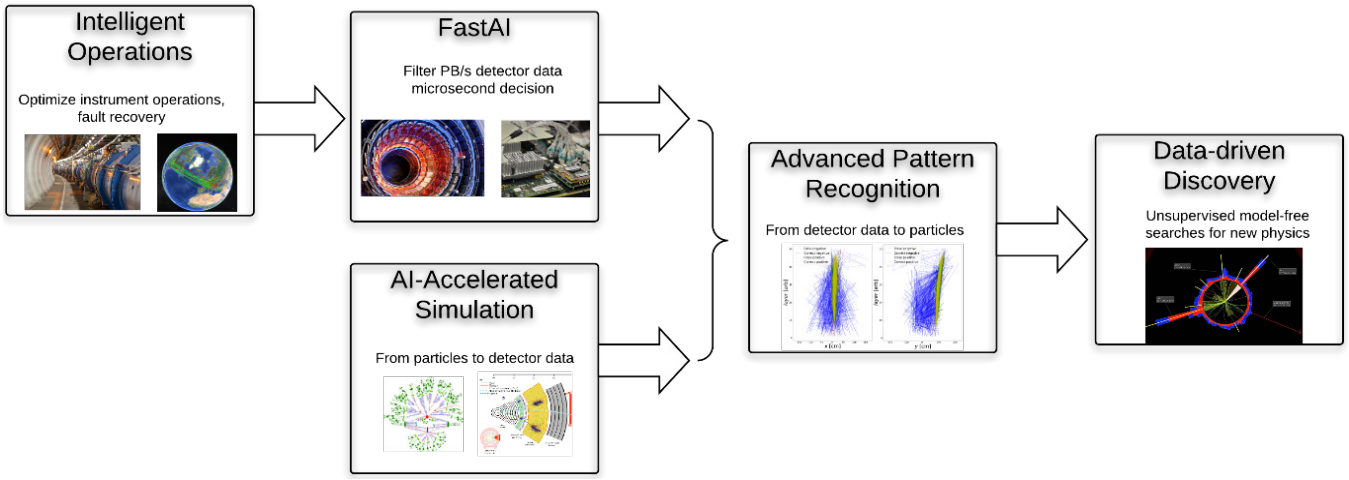


Figure 4.3 *AI-enabled ultra-fast event processing chain for HEP experiments.*

potential problems (see Chapter 10, AI Foundations and Open Problems). These techniques will allow for optimization of the operations of these complex systems to prevent or mitigate the impact of certain faults and to accelerate the return to normal operations if a fault occurs, increasing the instrument science output.

ii) ML inference with microsecond-latency in particle physics trigger applications. At HL-LHC, each detector will produce petabytes of detector data per second. The experiments will rely on a "trigger" system built from custom hardware, plus FPGAs, CPUs, and GPUs processors to reduce these data rates to a more manageable 10GB/s. The first level of this trigger system will reduce the detector data rate by three orders of magnitude in 10 microseconds or less. The challenge is to do that without throwing away any collision event resulting from rare or new physics processes. AI advances will allow us to detect and preserve these precious events that would otherwise be lost forever, while still meeting the stringent data rejection and latency requirements. Advances in AI model architectures and in the use of inference hardware (e.g., FPGAs) will be needed (see Chapter 13, Hardware Architectures).

iii) AI-enabled, ultra-fast event processing chain. Over the next decade, accelerator facilities—such as the High-Luminosity Large Hadron Collider (HL-LHC), and the Deep Underground Neutrino Experiment (DUNE)—will transform high energy physics. These facilities will be precise and powerful tools that will enable both the discovery of new particles and in-depth studies of known particles and fundamental interactions. They will produce hundreds of petabytes of raw data every year, and exabytes of simulated and secondary data streams. These data volumes will preclude straightforward extensions of current approaches for detector data analysis. Collider physics can be described as a massive inverse problem, requiring techniques from data merging, data visualization, and large-scale inference, first to "deconvolute" the detector signals from thousands of particles traversing it, and then to reconstruct the primary collision event from the particle measurements.

Key to the success of detector deconvolution and, in general, to the analysis of any particle detector dataset is the availability of accurate, high-statistics simulations of the detector response to particle traversal. Currently, high-accuracy detector simulation is performed using the Geant4 toolkit. As an example, the availability of datasets with trillions of simulated collision events could significantly increase the sensitivity of precision measurements in the Higgs and W boson sectors at the HL-LHC and help provide the first evidence for physics beyond the Standard Model. Simulating a collision event at HL-LHC can take up to

O(1Tflop) of computation and output O(1MB) of data. Producing and storing one trillion Geant4 collision events would be a challenge even on an exascale system. To achieve their physics goals, the next generation of facilities needs an ambitious R&D program in generative models with the goal to simulate the detector response for one collision (or interaction) event with O(1Gflop) or less, while maintaining an accuracy comparable to the one achievable using Geant4 (see Chapter 10, AI Foundations and Open Problems). Once this goal is achieved, the next challenge will be to integrate the AI-accelerated simulation into a fast *in situ* data processing chain of AI models for pattern recognition, particle classification, signal/background discrimination, anomaly detection, and model-free searches. Running this "fast chain" on massively parallel systems (see Chapter 16, Facilities Integration and AI Ecosystem) will be vital to maximizing the discovery potential of the next generation of particle-physics experiments.

3. Advances in the Next Decade

Major advances in experiments are expected in the next decade. The Cosmic Microwave Background Stage-4 (CMB-S4), DESI, and LSST surveys will be operating on the ground and Euclid, eROSITA, SPHEREx, and WFIRST will be sending data from space. HL-LHC and DUNE will be taking data by the middle of the decade.

The cosmological survey landscape in the coming decade will offer exciting challenges at the data analysis front. Interestingly (from the AI perspective), it is not only the increase in data size compared to contemporary surveys but also the increased complexity of the data due to enhanced resolution and depth of the telescopes. In particular, DOE is interested in extracting fundamental physics knowledge from cosmological surveys and answering questions about the nature of dark matter and dark energy, constraining the mass of neutrinos and the number of non-relativistic species, and investigating the physics of the very early universe. These questions have led to a rich observational program, currently focused in the optical and microwave bands. Specifically, during the next 10 to 15 years, data will be obtained and analyzed from the following DOE-supported surveys: DESI, LSST, and the CMB-S4 experiment. The data will provide many AI challenges, from understanding and reducing systematic errors to the determination of the most valuable follow-up observations, to image analyses, and to the inference of the cosmological parameters that describe the physics of our universe.

HL-LHC will be a major upgrade of the LHC and of its detectors. The experiments will observe at least 50 times more proton-proton collisions per unit time. The increased statistics will push the precision of most measurements of the property of the Higgs boson against the detectors' systematic accuracy. If the experiments can reduce their systematical errors, particularly by simulating the response of their detectors with high accuracy, the following may be enabled: (1) understand the nature of the Higgs boson (is it a fundamental particle or a composite?); (2) probe directly or indirectly the existence of new Beyond-Standard-Model particles and interactions, and (3) probe the existence of heavy, weakly interacting particles which may be dark matter constituents.

DUNE will study with unprecedented precision and accuracy the physics of neutrinos and offer new windows into the origin of the universe matter-antimatter asymmetry. The DUNE detector is also capable of studying neutrino bursts from exotic cosmic events, such as the formation of a black hole. DUNE may also be the first detector capable of observing the exceedingly rare decay of a proton, allowing it to constrain the energy scale at which the three gauge interactions are unified in a single theory.

The new HEP experimental facilities will be some of the world's largest sources of

high-quality scientific data. Exploitation and analysis of these data sets will greatly benefit from integration within the larger AI ecosystem consisting of DOE's high-performance computing and high-performance networking (HEP) facilities. The data itself will be generated within HEP facilities and instruments whose operation will also avail of a number of AI capabilities in the sphere of high-speed data classification, selection, and reduction, and in real-time control and optimization. In some contrast to the situation described in Chapters 14 (AI for Imaging) and 16 (Facilities Integration and AI Ecosystem), HEP data sets are already subject to well-defined quality standards and the field has a long history of established practice in large-scale data management and the exploitation of ML techniques. For these reasons, HEP facilities are in an excellent position to take immediate advantage of AI-enabled methodologies as they become available. Because large-scale HEP experiments have already built a sophisticated infrastructure for distributed data management and analysis based on a hierarchy of storage and analysis hubs and platforms, an exciting opportunity for greatly enhanced scientific returns exists in embedding this capability within DOE's broader HPC and HPN infrastructure via AI-enabled smart edge services to HPC systems and AI-enhanced "just-in-time" HPN-based data delivery systems.

4. Accelerating Development

The amount and complexity of the next-generation data stream will require a concerted effort to combine new analysis and modeling and simulation methods that effectively leverage AI technologies. New cosmological surveys are rapidly coming online. Given the aim of these surveys to deliver cosmological parameter constraints at percent level accuracy, AI will play a central role in the analysis and interpretation of the data. The cosmology community has embraced this opportunity fully and is developing approaches for numerous tasks already. In the near future, the focus will be to establish the reliability and robustness of AI-based methods for the different application areas (see Chapter 10, AI Foundations and Open Problems). In particular, whenever high precision is required, it has to be ensured that the AI approaches do not lead to undesirable biases due to, for example, misclassification of objects. With LSST and DESI coming on-line very soon, many new approaches will be applied and tested. In particular, for LSST's Dark Energy Science Collaboration (DESC), simulated data challenges are being created that will provide excellent testbeds for many of these projects. Therefore, the highest priority in cosmology over the next few years will be to develop a roadmap that clearly establishes the best application areas for AI and a solid understanding of their error properties. Based on the findings, the cosmology community will be able to fully integrate AI in their data analysis approaches and pipelines and then take the next major steps as outlined in the Grand Challenge problems to create a well-integrated overarching approach for use of AI to enable major advances in cosmological inference, extract the maximum possible information from the data, and inform and optimize new observational strategies.

For AI to play a critical role in DUNE and HL-LHC data simulation, processing, and analyses, new AI models are needed which are well suited to the sparse, high-precision nature of the measurements from most HEP detectors. Pattern recognition algorithms developed for a 10 MPixel photo camera do not work out of the box for a detector with 10 million active pixels, which are millimeters or even meters apart. Likewise, AI image-generation techniques commonly used to simulate new images from a library of existing ones, do not meet the stringent accuracy requirements of HEP detector simulation, particularly when it comes to simulating the tails of detector response. In general, for AI to address HEP data challenges of the next decade, we will need to identify resource-critical applications (such as detector

simulation) and to develop ML models that are good at simulating and detecting extremely rare phenomena[1] with high efficiency and accuracy. Besides supporting a robust R&D program targeting select HEP grand challenges, the development needed to meet these grand challenges includes:

(1) Create usable tools for large-scale distributed training and optimization of ML models to enable physicists to scale up the complexity of their models orders of magnitude above the current "laptop-size."

(2) Develop training methodologies that are able to detect rare features in high-dimensional spaces while being robust against systematic effects.

(3) Design tools to quantify the impact of systematic effects of the accuracy and stability of complex ML models.

5. Expected Outcomes

The primary aim of ongoing and upcoming cosmological experiments is to further our understanding of the dark universe (dark matter and dark energy), the very early moments of cosmic evolution (inflation), and the make-up of the universe. These are profound questions in the area of fundamental physics. AI will enable the exploration of the data from the next-generation surveys in new and unexpected ways. The data amount and the complexity of the data will increase immensely in the coming years, and in some areas, traditional methods will break down due to the sheer data volume (e.g., no human will be able to look at each image taken by the new surveys). The ability to make a movie of the universe from its earliest moments until today and into the future will have a profound impact—AI in cosmology will be central to our quest to understand the universe in which we live.

AI algorithms will play a vital role in the next generation of particle physics detectors and accelerators from Intelligent Operations, to Fast AI for data selection. Simulating and processing the DUNE and the LHC detector data with high statistics and high accuracy will usher in a new era of precision physics at the Energy and Intensity frontiers that may shed light on HEP fundamental questions such as the scale at which nature's fundamental forces are unified, the origin of the universe's matter-antimatter asymmetry, and the constituents of dark matter. The introduction of model-free, unsupervised AI searches will further push the potential of discoveries that may transform our understanding of fundamental physics over the next decade.

6. References

1. Rosner, J., et al., Planning the Future of US Particle Physics, *arXiv:1401.6075*
2. Cavuoti, S., et al., Machine-learning-based photometric redshifts for galaxies of the ESO Kilo-Degree Survey data release 2, *MNRAS 452, 3100 (2015).*
3. Kremer, J., et al., Big Universe, Big Data: Machine Learning and Image Analysis for Astronomy, *IEEE Intelligent Systems* 32, 16 (2017).
4. Higson, E., Handley, W., Hobson, M., and Lasenby, A., Bayesian sparse reconstruction: a brute force approach to astronomical imaging and machine learning, *MNRAS 483, 4828 (2019).*
5. Lanusse, F., et al., CMU DeepLens: deep learning for automatic image-based galaxy-galaxy strong lens finding, *MNRAS 473, 3895 (2018).*
6. Krause, E. and Eifler, T., CosmoLike – Cosmological Likelihood Analyses for Photometric Galaxy Surveys, *MNRAS, 470, 2100 (2017).*
7. Heitmann, K. et al., Cosmic Calibration, *Astrophys. J., 646, L1 (2006).*

[1] For example, DUNE expects to observe O(1) proton decay candidate per year while processing O(PB/s) of detector raw data.

8. Albertsson, K., et al., Machine Learning in High Energy Physics Community White Paper, *arXiv:1807.02876*
9. Radovic, A., et al., Machine learning at the energy and intensity frontiers of particle physics, *Nature 560, 41 (2018).*
10. Ilten, P., Williams, M., Yang, Y., Event generator tuning using bayesian optimization, *JINST 12.04 (2017).*
11. Albrect, J., HEP Community White Paper on Software trigger and event reconstruction, *arXiv: 1802.08638*
12. Collins, J. H., et al., Extending the Bump Hunt with Machine Learning, *arXiv: 1902.02634*
13. Ball, R.D., et al., Parton distributions for the LHC Run II, *JHEP 04, 40 (2015).*
14. Aurisano, A., et al., A Convolutional Neural Network Neutrino Event Classifier, *JINST 11.09 (2016).*
15. Acciarri, R., et al., Convolutional neural networks applied to neutrino events in a liquid argon time projection chamber, *JINST, 12.03 (2017).*
16. Akiyama, K., et al. (Event Horizon Telescope Collaboration), First M87 Event Horizon Telescope Results. I. The Shadow of the Supermassive Black Hole, *Astrophys. J. 875, L1 (2019).*

This page intentionally blank.

05. Nuclear Physics

The nature of matter is the fundamental question in nuclear physics: what are the basic components of matter and how do they interact to form the elements that make up our universe? This question is not limited to familiar forms of matter, but also includes exotic forms, such as those that existed in the first moments after the Big Bang, and those that exist today inside neutron stars. In addition to the fundamental questions of how and why matter takes on specific forms, it is also important to understand how that knowledge can benefit society in the areas of medicine, nuclear energy, and national security. Nuclear experiments include a range of devices, from small- and intermediate-scale devices to very large detector programs at accelerator laboratories like the Relativistic Heavy Ion Collider (RHIC) at Brookhaven National Laboratory (BNL), the Continuous Electron Beam Accelerator Facility at Thomas Jefferson National Accelerator Facility (Jefferson Lab), and the Argonne Tandem Linac Accelerator System at Argonne National Laboratory (Argonne). Nuclear physicists also lead experiments at other user facilities such as the Large Hadron Collider (LHC) at the European Organization for Nuclear Research (CERN) (Figure 5.1), the Japan Proton Accelerator Research Complex (J-PARC), and the Spallation Neutron Source (SNS) at Oak Ridge National Laboratory (ORNL).

Nuclear theory is concerned with how quarks and gluons interact to form protons, neutrons, and other hadrons, as well as how those hadrons interact to form and determine the behavior of atomic nuclei. Studies of the formation and characteristics of nuclear matter in stellar explosions (i.e., supernovae) and neutron stars are among the most computationally intensive investigations currently underway.

The Nuclear Physics Long Range Plan identifies the priorities for the field. These are:

- Utilize investments in accelerators, detectors, and computational infrastructure.
- Develop a U.S.-led, ton-scale neutrinoless double beta decay experiment.
- An electron-ion collider is the highest priority for new facility construction.
- Invest in small and mid-scale projects and initiatives enabling forefront research, including theory.

Applications of nuclear physics for societal benefit are also important.

The multiscale, highly correlated, and high-dimensionality nature of the physics of the nuclear force leads to a rich set of phenomena in nuclear physics. AI techniques offer the possibility of increasing our understanding of this physics and making new discoveries, through a number of applications, detailed here.

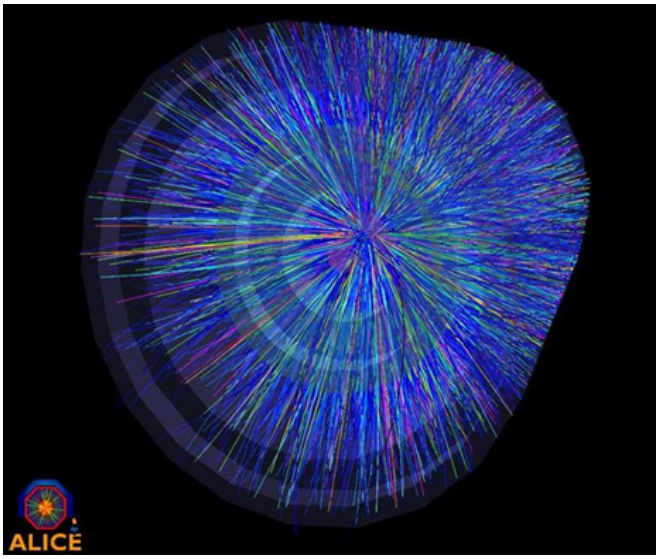


Figure 5.1 *An event display shows particle tracks from a lead-on-lead collision in the ALICE detector. Image courtesy of CERN, ALICE Collaboration [taken from https://www.energy.gov/science/np/articles/explaining-light-nuclei-production-heavy-ion-nuclear-collisions].*

1. State of the Art

Increasing data volumes from nuclear experiments and simulations have already led to a variety of AI approaches being employed in the field. These span nuclear theory, experiments at various scales, accelerator optimization and controls, and applied nuclear physics.

Nuclear binding energy, for example, is an essential property for understanding the production of nuclear species in astrophysical events such as supernovae and neutron star mergers. Some relevant binding energies cannot be measured directly and rely on nuclear models. Supercomputer calculations based on fundamental theory provide our best predictions for these binding energies and other important nuclear properties, but to reach the needed precision, these calculations become very computationally expensive. A team led by researchers from Iowa State University and Lawrence Berkeley National Laboratory (LBNL) developed a DL approach using a neural network trained with state-of-the-art supercomputer calculations [5]. The trained network estimates binding energies and other properties with precision beyond expectations from the available calculations. The researchers validated their approach by demonstrating consistency with available analytic and phenomenological extrapolation tools.

Experimental groups in all areas of nuclear physics are using AI techniques to characterize features in their data more quickly, efficiently, and with increasing sensitivity.

Experimental nuclear astrophysicists use the MUlti-Sampling Ionization Chamber (MUSIC) detector at Argonne to study the fusion of nuclei in stars and to understand explosive stellar phenomena such as Type I X-ray bursts and superbursts. Standard data analysis techniques require months to select relevant events. Data that were previously analyzed with standard techniques pass through algorithms based on the T-distributed stochastic neighbor embedding (t-SNE) approach [6] for unsupervised DL. The t-SNE approach was able to find clusters in 35 dimensions corresponding to different detector signals, clearly delineating previously identified $^{17}F(\alpha,p)$ reactions, providing a proof-of-concept for use in other reactions.

Analysis of the very complex data sets from heavy ion collisions at RHIC and the LHC already benefits from AI. Deep neural networks can connect specific moments of the complex particle correlations inside jets of hadrons with properties of the quark gluon plasma produced in the collision—in ways not previously predictable [8].

The GlueX experiment at Jefferson Lab utilizes a high-intensity photon beam and a large-acceptance particle detector to search for exotic hadrons. Individual collisions are reconstructed from fine-grained detector systems. A key use case of ML at GlueX thus far is in filtering those events containing rare reactions. GlueX demonstrated that Boosted Decision Trees achieved the required performance [7]. Another recent development in GlueX is a system of data quality monitoring using ML to evaluate images of data quality histograms in real time to identify problematic regions of the detector during the experiment's operation.

It is now known that neutrinos have mass. However, it is not known whether the neutrino is a Dirac or Majorana particle (i.e., the neutrino and the antineutrino are the same particle). To answer this question, nuclear physicists search for the lepton-number violating process of neutrinoless double beta decay, wherein two neutrons in an atomic nucleus are transformed into two protons without the usual emission of two antineutrinos. In such searches, it is paramount to differentiate a very small signal from background events that occur at rates orders of

magnitude larger. The backgrounds are dominated by intrinsic radioactivity in the detector along with instrumental backgrounds. Current neutrinoless double-beta decay demonstrator experiments are exploring different techniques to classify and separate the two-beta-electron signal from other classes of events in detectors, including large-scale liquid scintillators, semiconductor ionization detectors, bolometers, and high pressure gaseous Xenon TPCs. Geometric patterns of fired photomultiplier tubes are examined, or the pulses from charge or phonon collection are used. Most developments were begun with decision tree techniques, as event classification is the primary goal. Now, experiments are implementing DNNs and other DL methods. Improvements to the sensitivity to neutrinoless double beta decay have been demonstrated through more effective identification of background events.

As an example, the high spatial resolution of Xenon TPCs offer an additional handle for neutrinoless double beta decay searches beyond the excellent energy resolution at the 0νββ region of interest. The addition of spatial tracking information provided by the detector offers a topological separation of 2-electron events (resulting from a neutrinoless double beta decay event) from a single electron event of the same energy resulting from a background event. Calorimetric resolution of the Bragg peak of stopping particles differentiates between the start and end point of an electron track, and the topological signature (one Bragg peak or two Bragg Peaks) differentiates between single and double electron events (Figure 5.2). Topological signatures are important for reducing background rates and reaching the experimental sensitivity needed to learn the nature of the neutrino.

Spatially sparse image data, such as that found in high pressure Xenon TPCs, naturally lends itself to the application of CNNs for topological discrimination. ML, including DL, methods have shown excellent promise in this task of resolving signal and background events at the same energy in high pressure Xenon TPCs simulation and data, a neutrinoless double beta decay prototype, through the use of CNNs in three dimensions. Additionally, these networks were trained using scalable distributed learning techniques with spatially sparse convolutional networks and achieved the state of the art in less than 30 minutes of computational time.

Accelerator facilities are improving operations using AI technologies. At RHIC, efforts are under way to implement anomaly detection in the controls and AI methods in data mining. In addition, reinforcement learning is used along

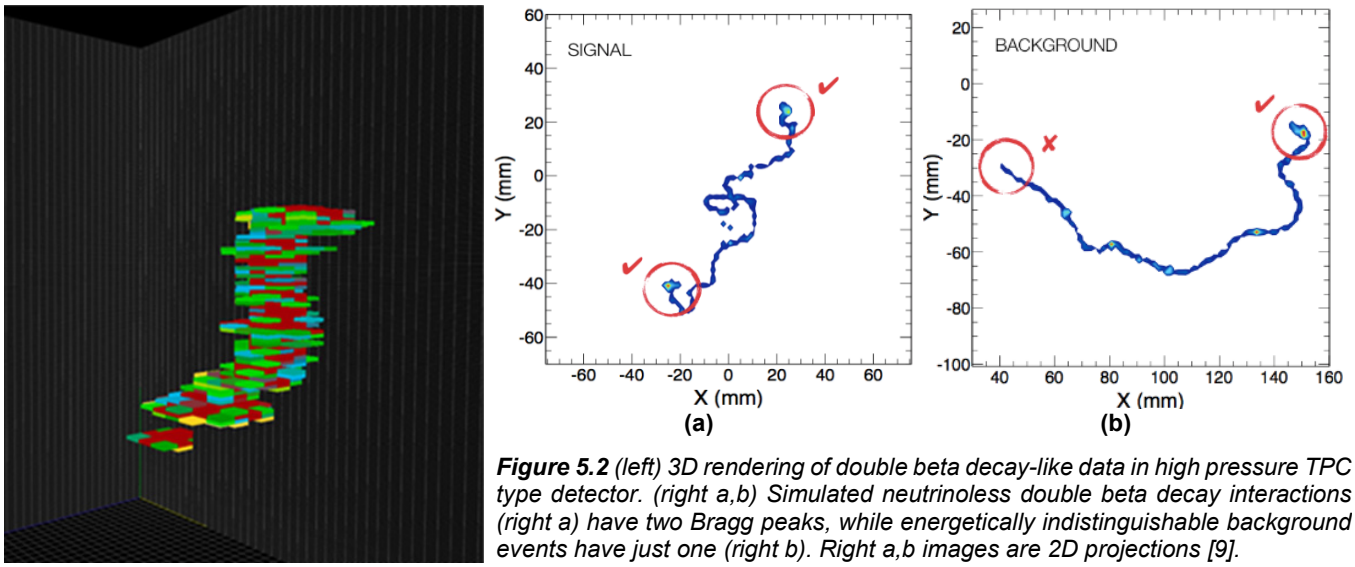


Figure 5.2 *(left) 3D rendering of double beta decay-like data in high pressure TPC type detector. (right a,b) Simulated neutrinoless double beta decay interactions (right a) have two Bragg peaks, while energetically indistinguishable background events have just one (right b). Right a,b images are 2D projections [9].*

with game theory to analyze client activities in the RHIC control systems [4]. Prognostics and errant beam prevention are becoming increasingly important in an age where we have many superconducting accelerators (superconducting magnets and super-conducting radio frequency), with high repetition rates and high power, and complex sensitive components. There is a much greater need for improved prognostics to avoid faults and to improve on recovery from faults. Many groups have efforts focusing on these areas, including improving mining large repositories of accelerator engineering data and introducing methods for real-time anomaly detection in operating systems.

An ongoing project at Jefferson Lab leverages ML to automate cavity trip classification. Traditional methods have been effective at identifying superconducting radiofrequency (SRF) trip causes, but are labor intensive and generate results in an asynchronous fashion. Identifying and correcting faults in real-time will have numerous benefits including improving the stability of the SRF system, providing a more reliable and available accelerator, and extending the energy reach. It will also provide important statistics and insights on cryomodule operations to engineering and SRF R&D staff while freeing them to focus on the future design and fabrication of SRF cryomodules. The project established a prototype system that reads data from the control system as faults occur, classifies it with a trained ML model, and outputs the result to subject matter experts. The system provides a cavity trip type, identifies the cavity causing the instability, and, potentially, can predict a trip before it occurs. It is a first step towards a diagnostic tool for daily use by operators to accurately identify a cause of a trip and apply precise response measures, avoiding unnecessary gradient reduction [10,11].

2. Major (Grand) Challenges

Advances in the use of AI/ML/DL techniques in nuclear physics will be driven by the volume and complexity of new data—both from experimental facilities (as described above) and from theory and simulation. The ability to discern physical causality and discover new phenomena will require the application of new technologies to augment human understanding. We note several grand challenges for better understanding the nature of matter in this section.

Generate detailed tomography of the proton/nuclei. This 3D tomography of hadrons and nuclear structure is not directly accessible in experiments. Obtaining the quantities of interest, such as generalized and transverse momentum dependent parton distribution functions (Generalized Parton Distributions (GPDs) and Transverse Momentum Distributions (TMDs)), involves an inverse problem. This is because these objects are inferred from experimental data using theoretical frameworks such as quantum chromodynamics (QCD) factorization theorems (e.g., collinear factorization, TMD factorization). Such a procedure allows one to connect experimental data to quantum probability distributions that characterize hadron and nuclear structure and the emergence of hadrons in terms of quark and gluon degrees of freedom.

Existing techniques to extract probability distributions from data have primarily been used to obtain a 1D tomography of hadrons, provided by parton distribution and fragmentation functions. These techniques usually rely on Bayesian likelihood techniques and Monte Carlo sampling methods, which are coupled with suitable parametrizations for the distribution functions of interest (Figure 5.3).

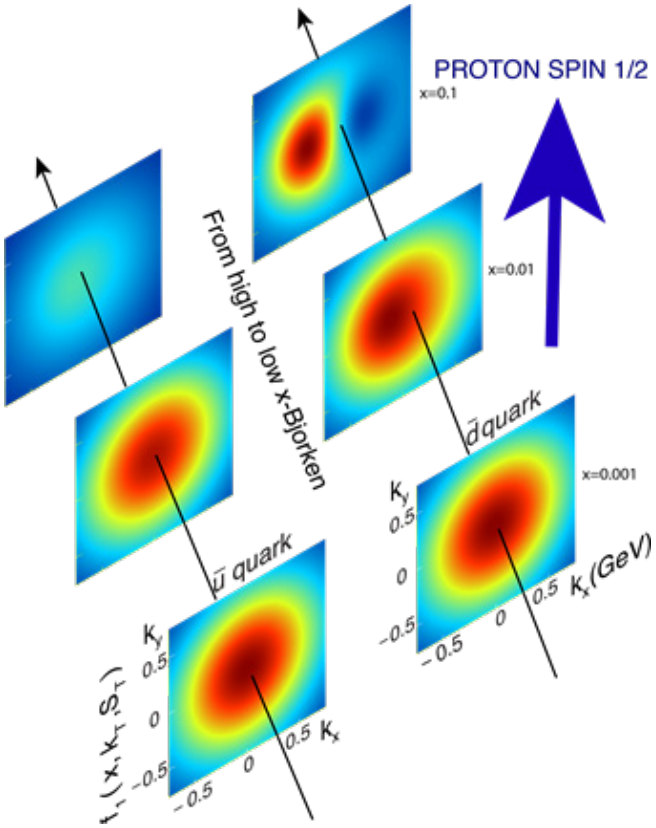


Figure 5.3 *A momentum space tomography of a hadron at difference slices in Bjorken x, for u and d anti-quarks. The images show how the variable x provides a filter to select different aspects of nucleon or nuclear partonic structure.*

In the Electron-Ion Collider (EIC) era, such methods need to be dramatically improved upon so that the full impact of the science can be assessed in real-time. This provides an important opportunity to utilize AI/ML techniques to obtain approximate solutions to the associated inverse problems. That is, to find an efficient mapping between the exabytes of experimental cross-section data and the theoretical objects of interest, namely the quantum probability distributions. Such a project will produce the next generation of QCD analysis tools that will provide rapid feedback between experimental data and a deeper understanding of strong interaction dynamics. Therefore, AI/ML methods will help guarantee maximum science output from the EIC.

Increase the understanding of matter/anti-matter in the universe. A better understanding of electroweak interactions are fundamental to understanding matter/antimatter asymmetry in the universe and neutrinoless double beta decay offers a window into these phenomena. CNNs offer the ability to reach beyond current technologies for neutrinoless double beta decay, thanks to the ability to quickly learn pattern recognition and discriminate important topological features. A significant challenge, however, will be validating a ML technique sufficiently well to ensure it performs on data in the energy region of interest.

With the availability of radioactive sources for calibration, such as Thallium in high pressure Xenon TPCs, researchers have access to a dataset with signal-like and background-like events that have a very similar topological signature to a neutrinoless double beta decay signal and background events, but at a different energy and with high statistics. The combination of available simulation, validation datasets, and very fast training times will allow experiments to perform an optimization campaign to build a robust neural architecture for fast analysis of neutrinoless double beta data with high confidence of similar performance on data and simulation. Additionally, the introduction of Generative Adversarial Networks (GANs) to model data/simulation discrepancies, with the ability to validate over large energy regimes, increases the confidence in a network trained on simulation + GAN datasets. The grand challenge in this space is to create an AI-centric workflow to distinguish neutrinoless double beta decay candidates from background, while using AI to validate simulations and ensure high-quality inference results on data.

Advance the understanding of nucleo-synthesis. Our understanding of nucleosynthesis is growing through studies of astronomical measurements, theoretical calculations, and experimental measurements of exotic nuclei generated at advanced experimental facilities.

Researchers are now working to extend deep learning to a wide range of important properties that govern the production of nuclei in the Cosmos. Further developments include applications to measure electromagnetic and weak transition rates in both stable and unstable nuclei. In addition, applications to improve scattering and reaction cross-sections based on fundamental theory appear feasible in light of the initial successes with binding energies. For example, incompletely converged supercomputer calculations of nucleon-nucleus cross-sections based on microscopic theory have appeared recently and, as with the binding energy example, a DL approach could extend those results to produce cross-sections at convergence with quantified uncertainties.

Nuclear astrophysics simulations—including core-collapse supernovae, X-ray bursts, and neutron star mergers—continue an inexorable march towards higher computational intensity, as increased physical fidelity is realized using higher spatial resolutions, longer physical times, and more complete microphysical descriptions. Anomaly detection for these very expensive (i.e., of order tens of millions of LCF node-hours) calculations becomes essential to ensure that scarce computational resources are not consumed in error. In addition, many of the requisite microphysics in these simulations (e.g., neutrino-matter interaction rates, thermonuclear reaction rates, and high-density equations of state) are recovered via the use of high-dimensional interpolation tables. ML techniques such as Gaussian process models and deep neural networks can replace traditional interpolation techniques while providing superior robustness.

When completed in 2022, the Facility for Rare Isotope Beams (FRIB) will be the world's most powerful rare isotope research laboratory. By producing intense beams of nearly 80 percent of the predicted isotopes for elements up to uranium, FRIB will enable researchers to make major advances in the structure, stability, and limits of nuclear matter, as well as in their interactions and decays (Figure 5.4). We anticipate that a variety of AI/ML approaches will be developed to address specific needs at FRIB, including beam generation, event characterization, detector response, experiment optimization, and data analysis.

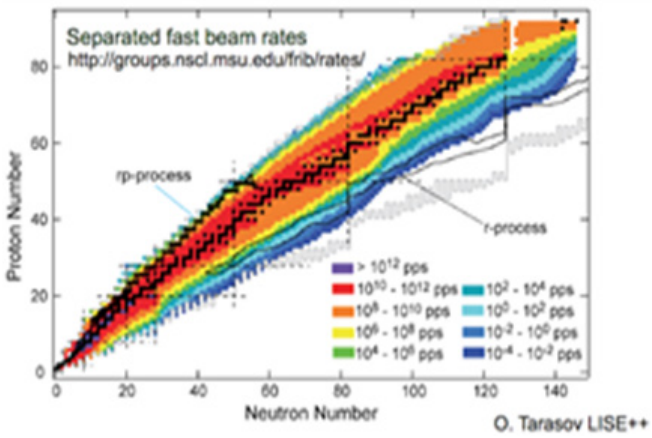


Figure 5.4 *The Facility for Rare Isotope Beams (FRIB) will provide unparalleled beam intensities of the most exotic nuclei.*

Transform the operation of accelerators and detector systems. In data analysis, experimental design and optimization, and even facility operation, AI/ML may provide approaches that are complementary to and offer improvement over traditional techniques. AI/ML studies can offer transformative progress in optimal operations of accelerators. In addition to the ongoing work at BNL and Jefferson Lab, FRIB operations will surely benefit. Production of high-purity, high-intensity beams of unstable nuclei and delivery with high efficiency to the FRIB experimental end stations present a daunting challenge. As data-taking runs for each measurement can be short, tuning time is important.

Time-consuming, multi-step beam generation efforts potentially limit the overall scientific productivity of the facility, as will the need to (on occasion) use sub-optimal beams with lower intensity. By utilizing supervised ML methods or reinforcement learning, it is anticipated that beam generation times can be significantly reduced compared to manual efforts, while simultaneously improving the quality of beams delivered to the end stations.

Detector systems used in nuclear physics experiments and nuclear physics applications will continue to generate higher fidelity data, which will drive needs for better data analysis methods, and, in some cases, for faster and high-fidelity edge-driven analysis.

AI techniques are being developed for event characterization, particle and photon tracking, particle identification, and energy reconstruction. Reconstruction of tracks in time projection chambers could be greatly improved with such approaches. At FRIB, logistic regression, fully connected neural networks, convolutional neural networks, and other approaches are being explored to identify event tracks in the Active Target Time Projection Chamber (AT-TPC). This step could be decoupled from fitting the tracks to determine reaction kinematics.

The enormous particle multiplicities in TPC's at heavy ion colliders cause track reconstruction to be slow and require complex correction for distortions due to the large charge load in the TPC. Application of ML to this problem would greatly simplify calibration and track reconstruction.

Methods to improving particle tracking through sophisticated magnetic spectrometers are also being developed through AI/ML. While the exact technique differs for different magnet configurations, room for improvement exists at all DOE Nuclear Physics (NP) accelerator facilities. At FRIB, correlating signals in the focal plane detectors of the magnetic spectrometers using a series of masks at the target location could be used to train corrections for offsets in initial particle angle and position. This could markedly improve the energy/momentum resolution of the focal plane spectra.

DNNs are being applied to complement existing Monte Carlo approaches for particle identification. Event shapes in multi-dimensional (detector signal) space can be used to train ML algorithms to recognize the location of foreground events in the presence of significant backgrounds. In calorimeters, DDN's allow sophisticated analysis of shower shapes to separate single photons, hadrons, and their decays.

Many modern detectors digitize the signals (waveforms) from each event. For example, new large-volume germanium detectors for gamma-ray spectroscopy will enable position sensitivity, i.e., determining not only the total energy deposited via gamma rays, but also the energy and position of the individual interactions within the detector. Spatial resolutions of a few millimeters will be possible, enabling so-called gamma-ray tracking, another area where ML is applicable. Gamma-ray tracking is the core operating principle of the Gamma-Ray Energy Tracking Array (GRETA) spectrometer, and AI/ML methods may transform current approaches using deterministic and probabilistic methods to reconstruct the path of multiple gamma rays from measured interaction positions and corresponding deposited energies. ML algorithms could be trained on the pattern of interaction points and energies with no assumptions of the underlying scattering processes. By focusing on differentiating events that are completely absorbed versus those that are partially absorbed, significant improvements are anticipated in determination of the peak-to-total, Doppler correction, angular distributions, and linear polarizations of events in GRETA. Improving the determination of gamma-ray transport parameters and transfer functions will improve the position resolution of the detector, especially for lower energy interaction points. Among other more established approaches, the use of GANs [3] for the discovery of these transfer functions is an attractive avenue of investigation. These techniques will be applicable to other detector systems, as well.

3. Advances in the Next Decade

The growth of AI techniques and the familiarization of nuclear physicists with those

techniques is anticipated to result in substantial advances in the next decade, which is particularly important given the planned increase in date volume and fidelity resulting from new experiments and facilities. In particular, the following advances are anticipated.

Extracting physics from simulations and other large-scale inverse problems. The coupling of higher-fidelity simulations that leverage HPC environments with the ability to conduct an ever-increasing number of simulations provides great opportunity to leverage AI to infer physics, manage and plan simulations, and tackle many other large-scale inverse problems, including 3D tomography, which relates to precision medicine (see Chapter 10, AI Foundations and Open Problems).

Data analysis. Data analysis methods will continue to advance the AI methods that are being leveraged for data analysis in both online and offline scenarios, where the online AI activities may be pushed closer to the sensor edge. Advances in particle tracking, particle identification, data fusion, and background reduction, as well as methods such as using shallow neural networks for curve fitting and other data analysis methods, will continue (see Chapter 10, AI Foundations and Open Problems).

Data management. Similar to data analysis methods, methods used to provide metadata, facilitate data discovery and data retrieval, and enable cross-experiment analyses will evolve thanks to AI methods that can reduce the now human-intensive task of curating data (see Chapter 12, Data Life Cycle and Infrastructure).

Facility operation. Experimental facilities are major investments in capital. Operating these facilities with minimal down-time and maximal user value provides the best return on investment and scientific outcomes. Improvements in beam diagnostics and control and beam-line planning will save human effort and produce better stewardship of the major investments (see Chapter 14 AI for Imaging).

Experimental design. More capable, AI-driven computing at the sensor edge will enable higher precision instruments to be developed and fielded at NP experiments. These advances may result in near-real-time tuning of detector parameters and better data acquisition decisions (see Chapter 15, AI at the Edge).

4. Accelerating Development

As outlined above, the sheer volume and complexity of nuclear physics data is increasing at a rapid pace. These increases are occurring across the enterprise of nuclear physics, from nuclear theory to experiment, and to the operation of facilities and the collection of data in support of nuclear science applications. Inference from these increasingly complex sources, and therefore physical understanding, is constrained even now by physicists' ability to examine, analyze, and interrogate data. The effective continued adoption of AI techniques into the nuclear physics workflow depends most critically on several factors:

- The development of AI/ML/DL techniques that are scalable from modest or scarce data volumes, to data volumes that can be exponentially larger (see Chapter 10, AI Foundations and Open Problems).
- AI approaches for anomaly detection and decision support that can be used in operating environments where expensive resources (e.g., accelerator beamlines and leadership-class supercomputers) are being used (see Chapter 15, AI at the Edge).
- Creation of new data analysis techniques for analyzing and interpreting the large multidimensional data sets produced by heterogeneous sensor networks, and methods of performing online sensor and sensor network reconfiguration to optimize performance. Two techniques of particular interest are the use of unsupervised learning

methods for the discovery of multi-dimensional patterns and the development of underlying models, as well as online learning techniques that are able to use streaming data to adapt to changing conditions across a network in real time (see Chapter 10, AI Foundations and Open Problems and Chapter 15, AI at the Edge).

- AI techniques that can optimize the design of complex, larger scale experiments could completely revolutionize the way experimental nuclear physics is done (see Chapter 10, AI Foundations and Open Problems).
- AI techniques can facilitate the collection and analysis of metadata, facilitating data reduction tasks to better document experimental conditions and better facilitate nuclear data evaluation and the 'interoperability' of data resulting from complex experiments (see Chapter 12, Data Life Cycle and Infrastructure).

5. Expected Outcomes

- One of the fundamental goals of nuclear physics is to understand how interactions between quarks and gluons ultimately manifest in the structure and binding of nucleons and nuclei. Approximate symmetries found in nuclear physics are thought to have origins not only in the underlying interaction, but also in the complicated many-body physics of the problems. AI has the potential to aid human understanding of these complex systems through improved methods that discern the origins of these symmetries and the emergent behavior that is often observed.
- Applications of AI in nuclear physics will produce a paradigm shift in the way information is gathered, stored, analyzed, and interpreted from the large amount of data obtained from scattering and decay experiments. With the aid of AI, experiments that require years of analysis will see decisions on optimization and results in near real-time. The accessibility of the data to the wider nuclear physics community would create a connectivity across experiments not seen before. This connectivity will become the standard rather than the exception in understanding nuclear phenomena from the laboratory to the universe.
- In a practical sense, radioactive and stable isotopes are critical to several societal needs. They are essential for energy exploration and innovation, medical applications, national security, and basic research. The utilization of AI to optimize the choice of reactor parameters, exposure time, and sample composition poses the potential to significantly increase the reliable and cost-effective production of isotopes, thereby impacting national needs in these areas.

6. References

1. Lee, I. Y., Gamma-ray tracking detectors. *Nucl. Instrum. Meth. A* **422**, 1-3 (1999), 195-200.
2. Deleplanque, M. A., et al., GRETA: utilizing new concepts in gamma-ray detection. *Nucl. Instrum. Meth. A* **430** 2-3 (1999), 292-310.
3. Goodfellow, I., et al., Generative Adversarial Nets. *Adv. Neural Inf. Process. Syst.* **27**, (2014) 2672–2680.
4. Gao, Y., J. Chen, T. Robertazzi, and K. A. Brown. Reinforcement learning based schemes to manage client activities in large distributed control systems. *Phys. Rev. Accel. Beams* **22**, 014601, January 2019.
5. Negoita, G. A., et al., Deep learning: Extrapolation tool for ab initio nuclear theory. *Phys. Rev. C* **99** (Oct. 2019).
6. Maaten, Laurens van der, and Geoffrey Hinton. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, Nov (2008): 2579-2605.

7. Dugger, M., et al., “A study of decays to strange final states with GlueX in Hall D using components of the BaBar DIRC,” arXiv:1408.0215 [physics.ins-det].
8. Lai, Y. S., arXiv:1810.00835.
9. Ferrario, P., et al., “Demonstration of the event identification capabilities of the NEXT-White detector,” arXiv:1905.13141 [physics.ins-det], accepted to JHEP 2019.
10. Solopova, A. D., et al., SRF Cavity Fault Classification Using Machine Learning at CEBAF. *Proc. 10th Int. Particle Accelerator Conf. (IPAC'19)*, Melbourne, Australia, May 2019, pp. 1167-1170.
11. Carpenter, A., et al., “Initial Implementation of a Machine Learning System for SRF Cavity Fault Classification at CEBAF.” 17th Int. Conf. on Accelerator and Large Experimental Physics Control Systems (ICALEPCS’19), New York, NY, USA, Oct. 2019, paper WEPHA025.

06. Fusion

The pursuit of fusion energy has required extensive experimental and theoretical science activities to develop the knowledge that will be needed to enable the design of successful fusion power plants. Even today, following decades of research in many key areas, including plasma physics and material science, much remains to be learned about the optimization of the tokamak—an experimental magnetic confinement machine that has the potential to produce controlled thermonuclear fusion power—or other paths toward fusion energy.

Data science methods from the fields of ML and AI offer opportunities for enabling or accelerating progress toward the realization of fusion energy by maximizing the amount and usefulness of information extracted from experimental and simulation output data (see Chapter 10, AI Foundations and Open Problems). While data-driven methods have long been used for specific roles in fusion research, such as real-time prediction of disruption risk in tokamaks [7], there is significant potential for impactful application to other areas, such as hypothesis generation and testing, optimization and acceleration of scientific workflows (Figure 6.1), boosting of experimental diagnostic data interpretability, model extraction and reduction, augmentation of plasma control effectiveness, and data-enhanced event and state prediction algorithms. The DOE recently sponsored workshops and assessments to determine optimal approaches and priority research opportunities for advancing fusion with ML and AI [2,8,18]. Grand challenges, anticipated advances, and expected outcomes discussed herein are abstracted from these and related assessment efforts.

1. State of the Art

ML models trained on large datasets have been employed in fusion energy research since the early 1990s. The most extensive applications have been to predict disruptions (catastrophic, sudden loss of plasma confinement due to growing instabilities) in tokamak magnetic confinement devices. Tokamaks have confined plasmas with temperatures in excess of 150 million degrees Celsius for many seconds and are the present leading candidate for a fusion power plant. There are a multitude of examples of ML application for disruption prediction, such as employing a neural network to predict high beta disruptions in real time from many axisymmetric-only input signals [21]; producing a multi-machine applicable disruption predictor for the Joint European Torus (JET) [20, 17] and the Axially Symmetric Divertor Experiment Upgrade (ASDEX-UG) [4,15,16]; demonstrating use of time series data and explicit look-ahead time windows for disruption predictability in Alcator C-Mod [5], DIII-D [11], and the Experimental Advanced Superconducting Tokomak (EAST) [13] (Figure 6.2); and demonstrating use of extensive profile measurements in multi-machine disruption prediction for the JET and DIII-D tokamaks with convolutional and recurrent neural networks [12]. Even with the growing use of ML methods

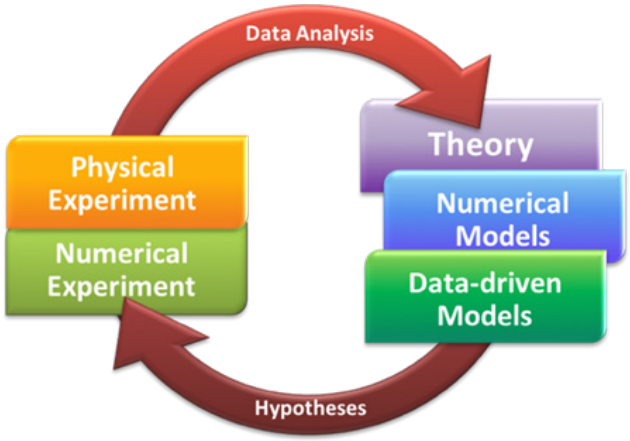


Figure 6.1 *Scientific discovery with ML includes approaches to bridging gaps in theoretical understanding through the identification of missing effects using large datasets, accelerating hypothesis generation and testing, and optimizing experimental planning to help speed progress in gaining new knowledge.*

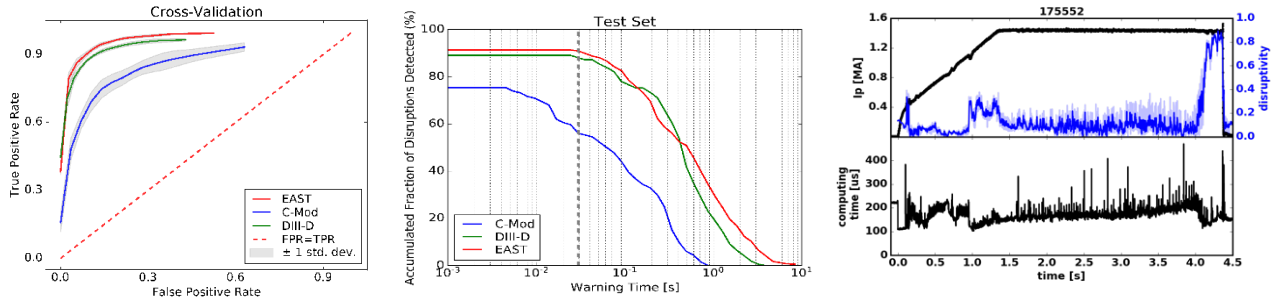


Figure 6.2 *The left two plots compare the performance of machine-specific disruption predictors on three different tokamaks (EAST, DIII-D, C-Mod). The rightmost plot shows the output of a real-time predictor installed in the DIII-D plasma control system, demonstrating an effective warning time of several hundred milliseconds before disruption [15].*

in fusion energy science applications, very little attention has been given to uncertainty quantification. Due to the inherent statistical nature of ML algorithms, comparing model predictions to data is nontrivial since uncertainty must be considered [19]. The predictive capabilities of a ML model are assessed using the model response as well as the uncertainty, and each aspect is critical to the combined effectiveness of real-time and offline applications.

In addition to the rapid growth in tokamak disruption predictors, in recent years applications of ML and statistical inference to fusion research have expanded to include model reduction for code acceleration [14], plasma control [6], and physics discovery [3,10].

2. Major (Grand) Challenges

The principal challenge in fusion energy research for the coming decades is to determine the key solutions that would establish the viability of a fusion power plant. The work on components of this overarching challenge is expected to grow, developing in perhaps unanticipated directions with the arrival of new burning plasma experiments such as ITER [1]. A recent joint Fusion Energy Sciences (FES)/Advanced Scientific Computing Research (ASCR)-sponsored workshop [2] identified a set of seven priority research opportunities for the application of ML to accelerate this process. These priorities were used to formulate the four Grand Challenges in this area.

Maximize predictive understanding of fusion plasmas and the burning plasma state. A central challenge for the advancement of fusion science toward the realization of fusion energy is the achievement of sufficiently predictive understanding of confined plasmas and, in particular, the burning plasma state. While both computational theoretical and experimental studies have produced substantial understanding of fundamental fusion plasma phenomena, significant progress is needed to enable high confidence design of operational power plants. For example, further understanding of energetic particle behavior in tokamak burning plasmas is needed to enable calculation of power plant performance and first wall impacts. Divertor function in self-heated tokamak plasmas must be projected to enable design of waste heat and exhaust handling solutions in a power plant. Much of this predictive understanding may still be undiscovered in data collected from fusion experiments and produced by simulations over the last ~50 years. Maximizing predictive understanding from data, both available and produced in the future, will be significantly aided by design and application of specialized ML methods.

This challenge can be addressed further through the development of specialized infrastructure, for which requirements are

tightly coupled to the unique nature of fusion experimental and computational resources. For example, neither experimental nor simulation data produced today are typically archived or made accessible in ways appropriate for large-scale application of ML methods. The Fusion Data Machine Learning Platform [2] is envisioned as a novel system for managing, formatting and curating fusion experimental and simulation data, with the goal of dramatically improving usability of data for ML algorithms. Such a platform is needed to enable unified management of both experimental and simulation workflows for ML, by supporting sufficiently rapid access to data from multiple experimental and computational sources (Figure 6.3). Fusion-specialized tools will be needed to enable efficient access to multi-machine and simulated data, either centralized or distributed, and to enable automated generation of fusion metadata for supervised learning.

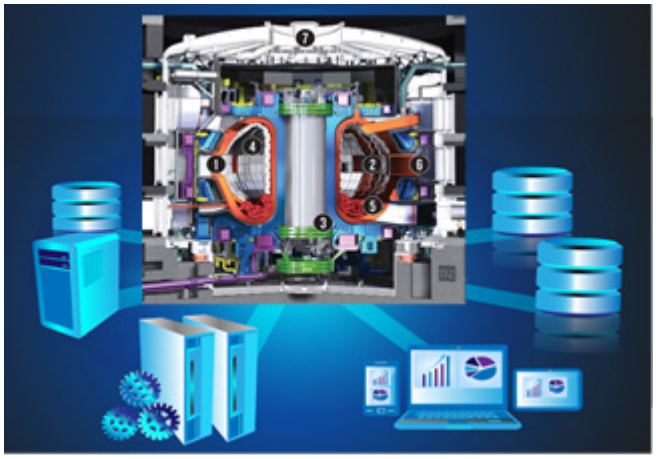


Figure 6.3 *Vision for a future Fusion Data Machine Learning Platform that connects tokamak experiments with an advanced storage and data streaming infrastructure that is immediately queryable and enables efficient processing by ML/AI algorithms.*

Key goals in this area for the next 10 to 15 years include the deployment of an effective Fusion Data Machine Learning Platform, characterized by extensive integration into the U.S. and international fusion workflow, and development of the relevant enabling algorithmic and computer science solutions specific to maximizing fusion plasma predictive understanding from plasma confinement experiments and simulations.

Enable real-time understanding in long-pulse tokamak experiments. The advent of long pulse, burning plasma, large-scale international fusion experimental devices will drive unique needs to extract the maximum amount of information from increasingly large and rapid real-time streams of data (Figure 6.4). These long pulse experimental devices will provide the first examples of the unique real-time data streaming and analysis requirements that will be posed by an operational fusion power plant.

Addressing this challenge will require interpreting and reducing fusion data at the source, as well as along the processing pipeline. The requirements for generation of real-time understanding and the nature of long pulse tokamak data streams are significantly unique to fusion experiments and burning plasma devices soon to be online. As such, they demand unique solutions and unique specific deployments of analysis systems. The effort will include integrating large numbers of fusion-specific data sources (multi-code, multi-machine, multi-diagnostic) to produce statistically supported interpretations, quantify uncertainties, and yield more understanding than the sum of individual sources. In particular, enabling federated, multi-institution collaborations on very large scales will pose unique problems. AI and ML methods are expected to be instrumental in addressing this challenge by providing methods for managing the increased data scales and unique fusion data types, as well as fusion-specific tools for enhancing interpretability.

Key goals in this area for the next 10 to 15 years include development of AI methods that will enable: a) *in situ*, in-memory analysis and reduction of extreme-scale simulation data as part of a federated, multi-institutional workflow, and b) ingestion into the new Fusion Data Machine Learning Platform and analysis of extreme-scale fusion experimental data

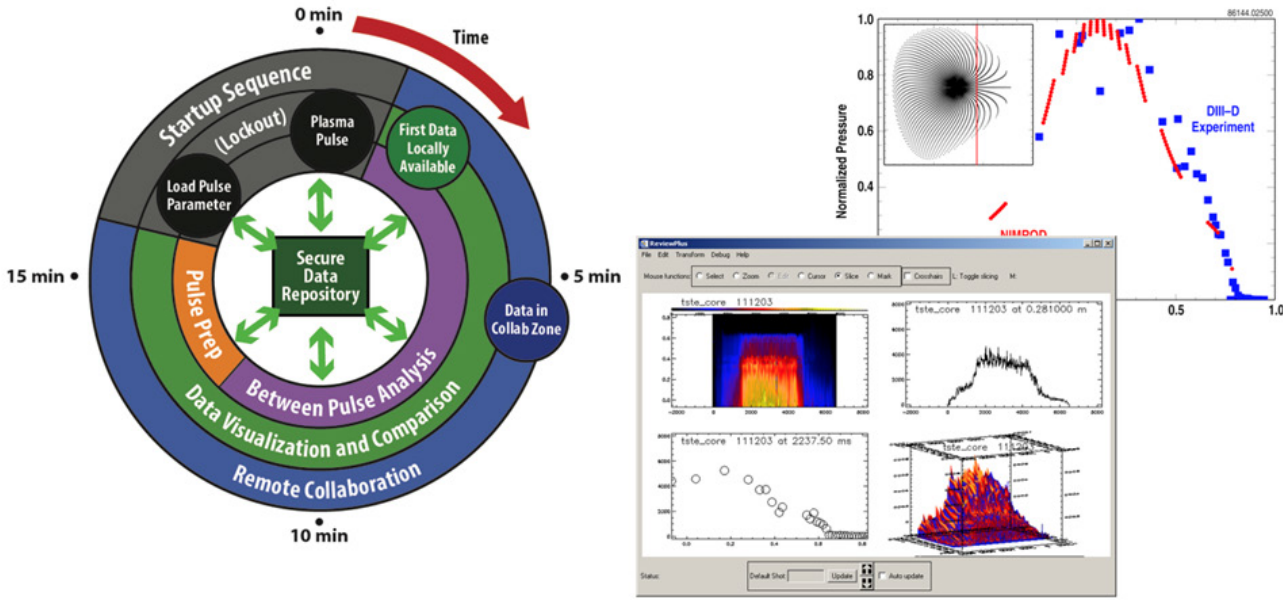


Figure 6.4 *The shot cycle in tokamak experiments includes many diagnostic data handling and analysis steps that could be enhanced or enabled by ML methods. These processes include interpretation of profile data, interpretation of fluctuation spectra, determination of particle and energy balances, and mapping of MHD stability throughout the discharge.*

for real- or near-real-time collaborative experimental research.

Develop models that bridge gaps in fusion plasma confinement and stability prediction. Fusion energy science is significantly challenged by existing gaps and uncertainties in the understanding of fusion-specific plasma physics, coupled with the increasing importance of simulations and analyses in closing these gaps. For example, while great strides have been made in modeling plasma phenomena that contribute to energy and particle transport in a tokamak, sufficient predictability has not been achieved, and the yet-unseen burning plasma regime is expected to yield further new phenomena that must be represented in models. Sufficient predictability of crucial performance-limiting and potentially disruptive instabilities such as tearing modes in tokamaks must also be achieved to enable operational scenarios and control for a reliable power plant.

ML offers techniques that can combine theoretical and data-driven models in hybrid systems that better represent the underlying dynamics specific to such fusion plasma phenomena. This approach has already been used successfully in fusion research [3,10], and is expected to play an increasingly important role in managing uncertainties and knowledge gaps in the coming era of long pulse burning plasma experiments.

Key goals in this area for the next 10 to 15 years include the development of interpretable ML methods and model extraction and reduction techniques that will help guide future experimental campaigns and help close gaps in the understanding of physics. Hybrid or other ML-informed models will be developed to enable sufficient predictability with quantified uncertainties for fusion plasma confinement, instabilities, plasma-wall interaction, and other critical physics areas.

Establish the plasma prediction and control solutions for sustained fusion power plant operation. A viable tokamak-based fusion power plant must have high-reliability, high-performance plasma control to ensure very low rates of operational interruption and system failure. Both control physics and control algorithm mathematics requirements for fusion plasma control are uniquely challenging due to their extreme nonlinearity, degree of multiphysics overlaps, resource limitations,

reliability requirements, and range of bandwidths involved. A key requirement is therefore to use data-driven methods to contribute to control-level modeling, management and interpretation of real-time data for control, optimal trajectory determination, and real-time prediction to support continuous and asynchronous actions and prevent faults (Figure 6.5).

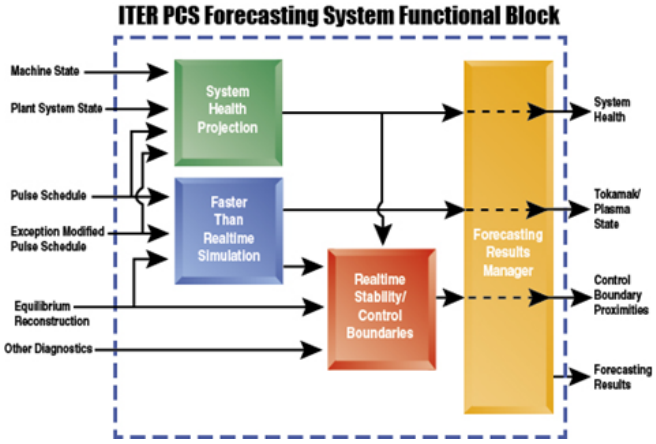


Figure 6.5 *The ITER Plasma Control System (PCS) Forecasting System will include functions to predict plasma evolution under planned control, plant system health and certain classes of impending faults, as well as real-time and projected plasma stability/controllability, including likelihood of pre-disruptive and disruptive conditions. Many or all of these functions will be aided or enabled by application of ML methods.*

Use of data-driven methods in control modeling and algorithm design always poses a challenge to operational application, due to the difficulty of quantifying uncertainty and reliability of performance with such approaches. The challenges specific to fusion are particularly demanding of advances in scientific understanding, as well as mathematical control theorems, due to the combination of multiphysics and range of bandwidth and plant integration scales. These characteristics dramatically amplify the fundamental challenge of operating the most complex, control-intensive power plant ever envisioned by mankind reliably for months at a time with extremely limited sensor and actuator resources (compared with present-day fusion devices).

Key goals in this area for the next 10 to 15 years include the identification of areas of fusion plasma control research that will most significantly benefit from ML/AI-augmented control algorithms, including data-driven methods that enable the prediction of key plasma phenomena and plant system states, allowing critical real-time and offline health monitoring and fault prediction. Mathematical approaches must be developed for quantifying the uncertainty of the data-driven fusion plasma models identified and the reliability of corresponding plasma control algorithms. Methods must be developed and qualified for extracting the required level of real-time control knowledge from limited diagnostics in a fusion power plant environment, while accomplishing the required level of control authority from limited actuators.

Addressing these four grand challenges for the application of statistical inference, AI, and ML methods to fusion research will contribute significantly to accelerating the development of solutions to many key problems on the path to fusion energy.

3. Advances in the Next Decade

Presently operating fusion experimental facilities will make significant advances in diagnostics, actuators, and accessible regimes in the coming decade, which will have equally significant impact on the data available for AI/ML applications (see Chapter 10, AI Foundations and Open Problems).

The advent of exascale high performance computing resources will provide a revolution in processing capabilities, enabling a similar leap forward in the effectiveness of large-scale data-driven algorithms (see Chapter 16, Facilities Integration and AI Ecosystem).

The most significant impact in the coming decade to the application of AI/ML methods to fusion problems is expected to be the availability of data from several key experimental facilities. ITER, the world's first

burning plasma experiment, will provide unique opportunities to study self-heated plasmas on a size and power scale relevant to a fusion power plant. JT-60SA [9], the largest long pulse superconducting tokamak in the world (until ITER operates), will explore advanced tokamak regimes not accessible by ITER. Data from these devices will provide extensive, novel groundwork for application of AI/ML techniques that maximize the information and understanding extracted. The amount and quality of these data will help better validate key components of plasma physics codes and reveal gaps in the understanding of the physics behind the models, thus suggesting improvements to the implementation of codes as well as the theory.

The deployment of a Fusion Machine Learning Data Platform could in itself prove a transformational advance, dramatically increasing the ability of fusion science, mathematics, and computer science communities to combine their areas of expertise in accelerating the solution of fusion energy problems.

4. Accelerating Development

The introduction of ML and AI into the scientific process for hypothesis generation and the design of experiments promises to significantly accelerate the scientific process by automating and accelerating the development of models and the testing of hypotheses (see Chapter 10, AI Foundations and Open Problems).

Perhaps the biggest obstacle in applying data science to hypothesis generation and experimental design is the availability of data and its lack of uniformity. In fusion, experimental data is limited by available diagnostics, experiments that cannot be reproduced at a sufficient frequency, and a lack of infrastructure and policies to easily share data. Furthermore, even with access to the existing data, there is still the obstacle that these data have not been properly curated for easy use by others. The Fusion Data Machine Learning Platform is envisioned as a step toward solving these problems (see Chapter 12, Data Life Cycle and Infrastructure).

Despite these gaps, we believe a research direction with the potentially highest payoff may be the integration of our knowledge of physics into ML models. Most existing AI/ML models are either purely data-driven or incorporate very simple physical laws and constraints. Without building the structure of physical laws into ML methods, it is difficult to interpret the predictions from data-driven models.

5. Expected Outcomes

Application of AI/ML methods to fusion energy research will accelerate progress toward realization of a commercial fusion power plant. It is very possible that the new capabilities offered will actually enable practical solution of problems not otherwise tractable even on a timescale of decades without use of data-driven methods.

Fusion energy offers an essentially infinite energy source with minimal environmental impacts, and high power density compatible with siting near high-demand population centers. Large-scale deployment of economically viable fusion power plants on worldwide grids have the potential to minimize the impacts of climate change and address the energy demands of the coming centuries. Fusion remains the only known energy option with virtually unlimited scalability to match growth in demand.

The long-term impacts of solving the relevant scientific challenges and achieving routine and widespread deployment of fusion power plants are potentially transformational, for society as a whole, and for the enterprise of science in particular.

6. References

1. Gribov, Y., et al., “ITER Physics Basis,” *Nuclear Fusion*, **47** (2007).
2. *Report of the Workshop on Advancing Fusion with Machine Learning April 30 – May 2, 2019.* https://science.osti.gov/-/media/fes/pdf/workshop-reports/FES_ASCR_Machine_Learning_Report.pdf
3. Baltz, E. A., et al., “Achievement of Sustained Net Plasma Heating in a Fusion Experiment with the Optometrist Algorithm,” *Nature Scientific Reports*, **7** (2017). doi:10.1038/s41598-017-06645-7
4. Bock, A., et al., “Advanced Tokamak Investigations in Full-Tungsten ASDEX Upgrade,” *Physics of Plasmas*, **25 (**2018).
5. Bonoli, P. T., et al., ”Lower Hybrid Current Drive Experiments on Alcator C-Mod: Comparison with Theory and Simulation,” *Physics of Plasmas*, **15** (2008).
6. Boyer, M. D., Kaye, S., Erickson, K. “Real-Time Capable Modeling of Neutral Beam Injection on NSTX-U Using Neural Networks,” *Nuclear Fusion*, **59** (2019).
7. Cannas, B., Cau, F., Fanni, A., Sonato, P., Zedda, M.K., and JET-EFDA Contributors, “Automatic Disruption Classification at JET: Comparison of Different Pattern Recognition Techniques,” *Nuclear Fusion*, **46** (2006).
8. Maingi, R., et al., “Summary of the FESAC Transformative Enabling Capabilities Panel Report,” *Fusion Science and Technology,* **75** (2019).
9. Giruzzi, G., et al., “Physics and Operation Oriented Activities in Preparation of the JT-60SA Tokamak Exploitation,” *Nuclear Fusion*, **57** (2017).
10. Gopalaswamy, V., et al., “Tripled Yield in Direct-Drive Laser Fusion through Statistical Modelling,” *Nature*, **565** (2019).
11. Hill, D.N., et al., “DIII-D Research Towards Resolving Key Issues for ITER and Steady State Tokamaks,” *Nuclear Fusion*, **53** (2013).
12. Kates-Harbeck, J., Svyatkovskiy, A., Tang, W., “Predicting Disruptive Instabilities in Controlled Fusion Plasmas Through Deep Learning,” *Nature*, **568** (2019).
13. Li, J., et al, “A Long-Pulse High Confinement Plasma Regime in the Experimental Advanced Superconducting Tokamak,” *Nature Physics,* **9** (2013).
14. Meneghini, O., et al., “Self-Consistent Core-Pedestal Transport Simulations With Neural Network Accelerated Models,” *Nuclear Fusion*, **57** (2017).
15. Montes, K. J., et al., “Machine Learning for Disruption Warning on Alcator C-Mod, DIII-D, and EAST,” *Nuclear Fusion*, **59** (2019).
16. Rea, C., et al., “Disruption Prediction Investigations using machine learning tools on DIII-D and Alcator C-Mod,” *Plasma Physics and Controlled Fusion*, **60** (2018).
17. Rebut, P-H., “The Joint European Torus (JET),” European Physical Journal, 43 (2018).
18. Baker, N., et al. *Workshop Report on Basic Research Needs for Scientific Machine Learning: Core Technologies for Artificial Intelligence*. doi:10.2172/1478744 (2019).
19. Smith, R. C., “Uncertainty Quantification: Theory, Implementation, and Applications,” SIAM, Philadelphia (2014)
20. Windsor, C. G., Pautasso, G., Tichmann, C., Buttery, R. J., Hender, T. C., JET EFDA Contributors and the ASDEX-UG team, “A Cross-Tokamak Neural Network Disruption Predictor for the JET and ASDEX Upgrade Tokamaks,” *Nuclear Fusion*, **45** (2005).
21. Wroblewski, D., Jahns, G. L., Leuer, J. A., “Tokamak Disruption Alarm Based on a Neural Network Model of the High-Beta Limit,” *Nuclear Fusion*, **37** (1997).

This page intentionally blank.

07. Engineering and Manufacturing

Over the last decade, advances in technologies, such as sensors, networks, and control systems, along with the rise of data analytics and artificial intelligence (AI) approaches, such as machine learning (ML), have led to increasing discussion of holistic approaches to manufacturing and engineering (see Chapter 15, AI at the Edge). Terms such as “smart manufacturing,” “the Internet of things,” and “digital twins” are used to refer to these types of transformational approaches, with the concept of optimization expanding to include an entire lifespan, from raw materials to shape/topology to manufacturing process to end use.

The future of manufacturing hinges on the ability to bring new ideas and custom products to market faster than ever before while reducing cost, energy use, and waste products. A major effort is under way to use distributed manufacturing and products designed for a circular economy to shrink the supply chain to the benefit of local communities. Obstacles include: disruptions in the supply chain due to natural disasters; changing economic costs (tariffs, transportation costs, etc.) or new regulations; inability to optimally utilize differing raw materials; appropriate data collection; weakness in altering processes in real time; and cybersecurity threats, among others. The goal is to overcome these obstacles in an optimal way to the benefit of the manufacturer, consumer, and environment.

1. State of the Art

AI has yet to have a major impact in manufacturing and engineering, but in the handful of examples provided here one can easily see the potential it has to change industry. To date, some of the initial forays into using AI have focused on smart manufacturing (improving efficiency and reducing waste), generative design, and autonomous robotic assembly.

Manufacturers of smaller batches, and ones who produce many different variants of similar designs for consumers who want a customized product, need robots on the assembly line to perform tasks autonomously rather than automatically. Typical automation is not profitable at this level, and this is often referred to as the “Batch Size 1” or “Order of One” problem. What is meant by “autonomous” is that the robots are not reprogrammed step-by-step to complete the new assembly; rather, they independently learn how to optimally assemble one variant or another. Basically, the robots are provided the fundamentals to learn how to assemble on their own.

Siemens Corporate Technology has managed to solve this problem for some simple assemblies [1]. They have done this by semantically converting the parts and process information into ontologies and knowledge graphs, thereby converting implicit information into explicit. Previously, the robots had to be taught through code, but now the robots analyze the CAD drawings and find the corresponding solution to assembly (Figure 7.1). An added benefit is that the robots are also able to correct some faults without having this option explicitly instructed beforehand. If a part slips and falls or is needed on the other side of the assembly, one robotic arm can stop and pick it up or pass it off to its partner and the assembly can continue on unimpeded.

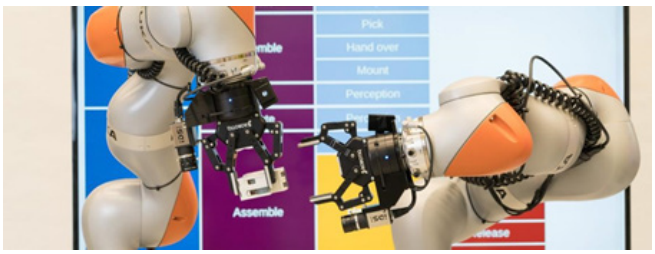


Figure 7.1 *This Siemens two-armed robot uses AI to interpret CAD instructions and assemble parts.*

In Korea, Siemens’ largest IT provider, LG CNS, works across a variety of industries using

its cloud-based smart factory service. This service helps manufacturers automate production and keep track of efficiency throughout the entire process. Collecting data over large swaths of production history and placing it into easily accessible databases, manufacturers have used Microsoft's *Azure Machine Learning* to predict defects before they happen [2]. While not perfect, it greatly minimizes costs due to delays on the production line and the subsequent waste these defects cause.

Generative design is one of the newest areas through which AI has had an impact on manufacturing. This is a two-step iterative process that, based on design goals, generates a number of possible outputs that meet the specified constraints. A designer then tunes variables in these outputs (over previously set minimal and maximal values) to reduce and optimize potential outputs that meet the aforementioned constraints. Generative adversarial networks are often used to drive the underlying optimal design. Airbus has employed this technique to improve the design of the partition that separates the passenger compartment from the galley in the Airbus A320 cabin (Figures 7.2 and 7.3). Their design goals focused on a reduction in weight with constraints in maximal width, strength to support two jump seats during takeoffs and landings, and number of airframe attachment points [3]. Note that the models used to assess designs are typically reduced-order surrogates, and the impact of their fidelity to optimality of the final design is an open question.

2. Major (Grand) Challenges

Since additive manufacturing (AM) is in relatively early stages of development, it can simultaneously gain the greatest benefit from advanced simulation, data analytics, and AI approaches and offers the greatest flexibility and research resources for implementing those ideas. So, although the spectrum of engineering and manufacturing processes that can be impacted is much broader than AM, it will serve as our exemplar.

AM is revolutionizing manufacturing, allowing construction of complex parts not readily fabricated by traditional techniques. In addition, AM offers the possibility of constructing "designer materials" by adjusting process control variables to achieve spatially varying physical properties. AM is a unique application area due to its strategic importance to both U.S. industry and federal agencies (DOE, NNSA, DOD, NASA). Although there has been significant interest and investment in AM, the fraction of this investment devoted to modeling and simulation—not to mention data analytics, ML, and AI—is relatively small.

Modeling of the AM process allows both prediction of how the AM process variables (the "machine knobs") impact the resulting material microstructure (the forward problem) and the ability to control the AM process to manufacture parts with desired properties (the optimization problem). Our grand challenges tackle some of the most pressing problems in these areas, and the ones with the most potential to accelerate manufacturing.

We would be remiss if we failed to reference reports from two workshops held by the National Academies of Sciences, Engineering, and Medicine. The first, held in 2016, was titled "Predictive Theoretical and Computational Approaches for Additive Manufacturing" [4]. The second, held in 2018, was titled "Data-Driven Modeling for Additive Manufacturing of Metals" [5] and is of particular relevance to the current topic.

Optimally solve the Batch Size 1 problem in additive manufacturing. The ability to quickly design a new product, optimally, without going through an expensive simulation (let alone trial and error), is the path to solving the "Batch Size 1" problem in AM. This can be carried out through the creation of a high-quality surrogate model.

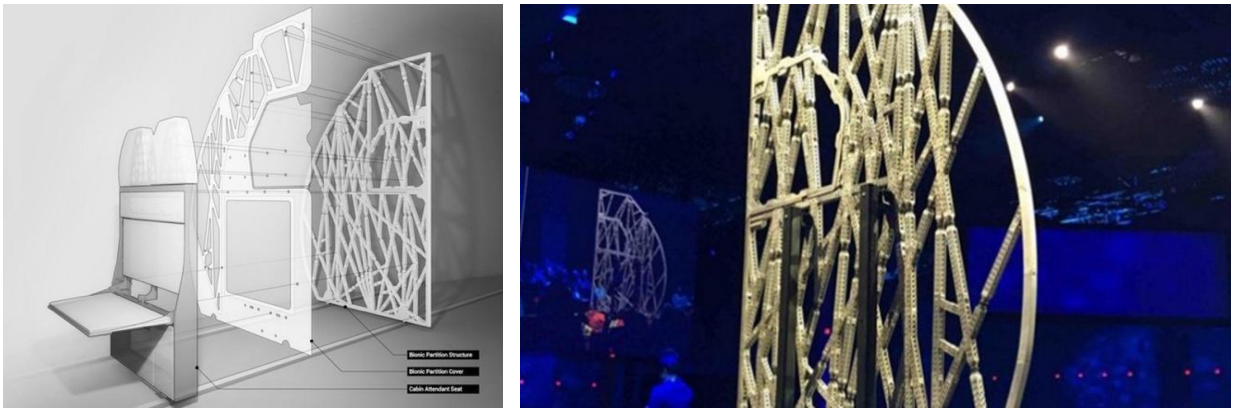


Figure 7.2 *This generative designed partition for the Airbus A320, with its seemingly random construction, has been optimally designed to be both lightweight and strong.*

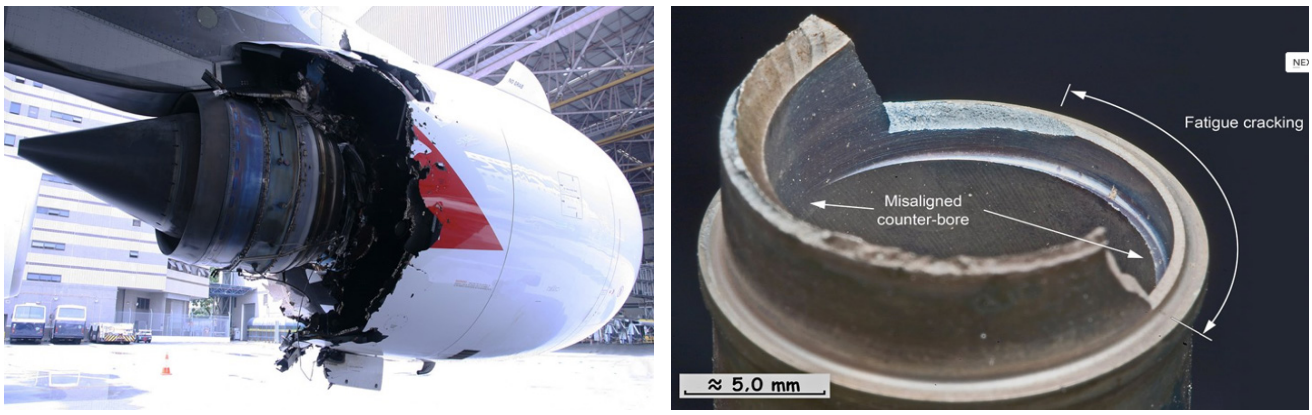


Figure 7.3 *Left: In 2010, an Airbus A380 sustained an uncontained engine rotor failure (UERF) of the No. 2 engine as it departed from Singapore while climbing through 7,000 ft. Debris from the UERF hit the aircraft, which led to significant structural damage. Bottom: The culprit was caused by metal fatigue in an oil feed stub pipe due to a slightly misaligned machining that left the pipe a little thinner on one side. [Australian Transport Safety Bureau Investigation #: A0-2010-089]*

Surrogate models (which includes reduced order models, see Chapter 10, AI Foundations and Open Problems) can play at least three roles in AM: (1) *a priori* optimization, (2) *in situ*, real-time process control, and (3) transferability of AI models between different devices and/or feedstocks—heterogeneous manufacturing. The first role encompasses both design and process optimization. Design optimization is the outermost loop and includes both shape and topology and implicitly local control of microstructure and properties. Although there has been significant research over the last few years in shape and topology optimization, it typically relies on extremely simplistic physical models. The extent to which model fidelity impacts "optimal" design is unknown. Similarly, process optimization (selection of parameters such as beam diameter, beam power, preheat, and scan strategy) also relies on approximate models. Improved surrogate models based on physics-informed AI models would enable more extensive exploration of both design and parameter space, ultimately accelerating qualification of AM parts.

The second role would have even more impact, but is also significantly more difficult. It requires access to the AM control system, an extensive array of sensors, and the ability to process data from the sensors during a build, analyze it in real time, and determine whether and how to alter any process control parameters. Since this would have to happen in a matter of seconds (between layers of a build), the data processing and analysis requirements are significant, and accurate, fast-running surrogate models are essential (see Chapter 15, AI at the Edge).

An additional twist to this entire process would be the ability to transfer a surrogate model trained on one AM system to another of a different design, or equally from one feedstock to another on the same AM system. Heterogenous manufacturing, where different systems build the same part to the same specifications, would open up the possibility of distributed manufacturing at an entirely new level. Here a CAD drawing for a given part, along with the system configuration and feedstock, would be used as inputs to the AI process to produce the optimal design for the system in question.

Couple material design with prediction and control of the additive manufacturing design process. Microstructures produced by AM processes are very different from those that arise from traditional manufacturing processes, such as casting and forging, and these differences can lead to extremely poor properties (strength, ductility, etc.) and unsatisfactory performance (likelihood of fatigue/cracking, lifetime, etc.). The good news is that the microstructures produced are strongly dependent on process parameters and conditions such as the beam power, scan speed, scan pattern, etc., and on geometry.

It should be noted that the microstructures produced by AM can be better or worse than those produced by traditional processes. Part of the difference is due to the unpredictability of response to process parameters. But another important factor is that most alloys were designed for traditional processes, and we cannot expect them to respond in the same way when manufactured using AM. Most of the alloys we use today were invented years or decades ago. The best high-performance aluminum alloy for pistons was invented after World War II and hasn't been improved upon for more than 70 years! We have to relearn how to innovate in metallurgy and manufacturing to be successful. Coupling an understanding of the fundamental materials science with prediction and control of the process dynamics through AI will allow us to design new materials and process characteristics to achieve desirable microstructures and properties (see Chapter 1, Chemistry, Materials, and Nanoscience).

Along this line, a very important component to manufacturing is a full understanding of part tolerances and lifetimes. Folding this into the entire aforementioned design process is a major goal in the future. Parts are precisely designed to meet specific tolerances and are engineered to function perfectly over a specified lifetime. Failure to meet these specifications can often produce catastrophic consequences. For these reasons, many potential applications of AI in manufacturing will need to be explainable (or interpretable). Parts designed by AI for use in the automotive, aircraft, or medical device industry (among many others) will need to have a full cost accounting of how they met the design specifications and to what tolerances.

Securely aggregate data across the manufacturing industry. A key challenge in maximizing our knowledge in the manufacturing process, and doing it robustly, is the collection of vast quantities of data across many different systems. This is hindered by the fact that companies do not want to share this data for fear of losing their IP and subsequent competitive advantage in the field. A path toward solving this challenge can be found in federated learning. This is an ML technique where the goal is to train a high-quality, centralized model in which the training data remains distributed over a large number of clients. For every iteration during the learning, each client independently computes an update to the current model based on its own data and then pushes this update to a central server, where it is aggregated to compute a new globally optimized model [6]. Through secure, federated learning, it is now possible with several AI applications to train the models without exposing the underlying data, even when attacked by a variety of adversaries [7].

Develop data and design tools for manufacturing in a circular economy. By 2050, the world's population will likely pass 10 billion. As the Earth's raw materials are not limitless, and global labor and the costs for these materials are on the rise, new solutions are needed to mitigate this emerging challenge. Circular economy business opportunities are one way manufacturing can grow and diversify under these pressures. In a circular economy, materials, and the resultant products, keep circulating in a high- value state of use, through supply chains, for as long as possible. The key challenge of transforming the current manufacturing ecosystem is providing design tools to develop products that are easy to remanufacture, recycle, or capture critical materials for reuse. Many locally sourced and sustainable materials are hard to integrate in product design and prototyping. The linear economy depends on mines and materials scale-up facilities across the globe for dependable supply. The challenging aspect of manufacturing for a circular economy is the ability to identify and optimize a supply chain using massive amounts of customer data on products currently in use, and incentivizing consumers to engage in the supply chain. New models need to be developed for tagging products nearing end-of-life and they need to use AI to optimize supply chain models to reduce fluctuations and disruptions.

Transition to smart engineering of products, services, and operations. Finally, engineering functions across the value chain need to evolve to integrate AI to create better products as well as the next wave of products. There are four major tasks here: (1) AI to reduce cost and accelerate time-to-market; (2) AI for optimization under uncertainty and constraints; (3) AI for real-time control and steering, and (4) AI for cradle-to-grave system state awareness.

The challenges here are large. From jet engines to consumer products, designers will need access to innovative tools and accessible computing services to partner with AI without a steep learning curve. The bootstrapping of industrial machines, with the ability to perform optimization under uncertainty and constraints, will require a range of new method development to respond to streaming data from machines and the fusion of complex datasets in real time. Here there is an additional need to provide robust safeguards such as physics constraints and strong cybersecurity. This can alter the trajectory of research in control systems for many domains where we need a cost-effective way to transform the legacy machines into smart machines. Ultimately, from large machines to connected products in the hands of consumers like smart phones, the cradle-to-grave awareness of the system state will create a new engineering ecosystem.

The four cases described above have commonality in the need for the integration of datasets across different engineering functions in an ecosystem of smart machines and smart products. This will enable lower cost, energy-efficient operations, and, ultimately, a sustainable products and consumption economy with responsible use of resources.

3. Advances in the Next Decade

Building an integrated software environment for manufacturing data and AI will provide researchers with the ability to merge data, ML models, computer vision, simulations, and knowledge to accelerate the state-of-the-art in manufacturing. Leadership from the national labs can improve the edge-to-exascale infrastructure needed to advance response time. For manufacturing, the ability to provide real-time quality control for products such as advanced batteries, complex parts, electronics, and sensors will be unique in the nation.

The future of manufacturing is closely linked to advances in intelligent cyber-physical systems for bringing new ideas and custom products to market faster than ever before. In addition, enabling creative but market-conscious design where one includes constraints based on cost, quality, lifetime, aesthetics, manufacturability,

recyclability, and supply-chain logistics will be critical for gaining an edge in manufacturing. Given the sheer complexity of this system, an AI-driven design process is a natural solution for this optimizing problem.

The response time and data needs can skyrocket if decisions are not made locally and models for the manufacturing process are not trained in smart ways. The need for software-defined sensors and edge computing is paramount for making progress in improving AI models for a digital and custom manufacturing future. It will also be likely that a new generation of low-cost imaging and metrology will need to be developed.

DOE recently launched the ReCell Center to capture critical materials such as cobalt from electric vehicle (EV) batteries. EV batteries that are 100% recyclable will be needed to meet the demand of a rapidly growing market, with few suppliers for critical materials for mobility. Application of similar concepts to consumer electronics can be transformational for design and reduction of e-waste and are estimated to unlock $90 billion economic value by 2030 [8,9]. AI can be used to connect manufacturing processes to adjust to changes in supply, develop intelligent process optimization to increase efficiency, and allow continuous improvement of products for a circular economy.

4. Accelerating Development

The main bottleneck in a competitive labor market is to develop a fully functional model for integrating AI all across the design, engineering, supply-chain management, resource planning, and manufacturing sectors. A good way to accelerate the transition is to allow development and testing of applications in a secured environment. The ideal framework will be to build pre-competitive tools and benchmarks for manufacturing with strong adherence to standards and safe keeping of proprietary data.

More research in secure, federated learning will be needed to not only accelerate AI's adoption by more users, but to know which AI methods can remain secure. A public-private partnership supporting the creation of such a hub for training and testing accessible AI tools will increase industry's access to the expertise in the national labs and academia. This will allow industries in the U.S., struggling to take advantage of AI for improving business efficiency for engineering services and manufacturing, to leverage these new capabilities and know that their data is secure. Another major challenge for applying AI to manufacturing is that the data is both noisy and expensive to collect. We have to efficiently design experiments to collect the most valuable data, design high-quality characterization and sensing modalities to get clean, pedigreed data, and then properly label that data. In addition, there is significant need for curated, publicly accessible datasets—both experimental and simulation. Efforts such as the NIST Additive Manufacturing Benchmark Test Series (AM-Bench) is a huge step in the right direction, but these efforts need more widespread support with an additional focus on community data formats.

Finally, AM systems themselves need to be more open. APIs for the control systems should be available to researchers to explore more advanced, real-time analysis, feedback, and process control, all of which is necessary before AI can be effectively deployed on these systems.

What must we do to accelerate development?

a) Automate the entire learning pipeline: The goal is for the machines to be intelligent and learn the production process with full awareness of the intent of the designer and certification/qualification goals (see Chapter 9, AI for Computer Science).

b) Determine the best AI techniques for developing surrogate models for the manufacturing process.

c) Combine sensor modalities and incorporate data from a fleet of machines instead of a single machine, as the data from a single machine is prone to variability and often provides poor statistics for creating an intelligent learning environment (see Chapter 10, AI Foundations and Open Problems).

d) Develop an open framework with standardization data formats and plug-and-play module capabilities, but designed with protection through secure, federated learning to make use of exact design specifications while maintaining proprietary knowledge.

e) Provide access to curated datasets so that we can accelerate reinforcement learning across the manufacturing industry.

f) Incorporate physics-informed AI to reduce simulation and data requirements in training (see Chapter 10, AI Foundations and Open Problems).

What are the top priorities?

- Couple current mod-sim efforts in manufacturing with a variety of AI-generated surrogate models to determine which are the most robust and least biased.
- Determine which AI techniques are amenable to secure, federated learning.
- Design experiments to determine which data is needed, and of what quality, in a few exemplar manufacturing processes to improve design optimization through AI.

How do we improve scale?

Several of the potential methods one might employ for the creation of surrogate models, for both design and optimization, will require a major scale-up effort. These simulations are already pushing toward the exascale level. Running hundreds or thousands of higher fidelity physics models to develop the appropriate training data, understand bounds, etc., will be a major challenge. Efforts to reduce this cost, or enable the transfer of trained models from one system setup to another, will likely be critical to the success of this effort.

Equally important will be the confrontation of data, which is currently limited in nature, quality, and size, to these models to constrain parameter space and perform effective optimization. Major efforts need to be made in increasing data collection, determining which data needs to be collected, and improving the quality of the data, as well as designing the data formats to be optimized for AI training.

5. Expected Outcomes

The domains of engineering and manufacturing span a large portion of the U.S. Gross Domestic Product (GDP) and investment in R&D. While U.S. consumption is high, much of the manufacturing is currently done outside the U.S. The supremacy of products and the ability to compete in a global marketplace can be accelerated through technological leadership and dominance in AI in manufacturing. Use of AI will become the primary way in which future workforces can participate in a distributed manufacturing ecosystem where design, supply-chain management, prototyping, and production will be managed by people with diverse skill sets, and distributed geographically, but they will be connected by a digital manufacturing backbone with strong integration of training data, accessible knowledge, and AI-enabled tools. This will allow entrepreneurs and small businesses to successfully compete on the world stage.

6. References

1. Zistl, S. “The Future of Manufacturing: Prototype Robot Solves Problems without Programming,” *Seimens.com Global Website.*

2. Microsoft UK Enterprise Team. “Better, faster, more efficient: AI meets manufacturing.” *Microsoft Industry Blog – United Kingdom* (6 June, 2018).
3. “Airbus: Reimagining the future of air travel.” *Autodesk Website*.
4. *U.S. National Committee on Theoretical and Applied Mechanics, Board on International Scientific Organizations, Policy and Global Affairs, and National Academies of Sciences, Engineering, and Medicine, Predictive Theoretical and Computational Approaches for Additive Manufacturing: Proceedings of a Workshop.* Washington, D.C.: National Academies Press, 2016. DOI: 10.17226/23646.
5. *Board on Mathematical Sciences and Analytics, National Materials and Manufacturing Board, Division on Engineering and Physical Sciences, and National Academies of Sciences, Engineering, and Medicine, Data-Driven Modeling for Additive Manufacturing of Metals: Proceedings of a Workshop.* Washington, D.C.: National Academies Press, 2019. DOI: 10.17226/25481.
6. Bonawitz, K., et al., *Practical Secure Aggregation for Privacy-Preserving Machine Learning.* Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. 1175-1191. Oct 30-Nov 3, Dallas, TX, 2017.
7. Kasiviswanathan, S. P., et al. What Can We Learn Privately? *The 49th Annual IEEE Symposium on Foundations of Computer Science.* 531-540. Oct. 25-28, Philadelphia, PA (2008).
8. Balde, C. P., et al. *The Global E-waste Monitor 2017: Quantities, Flows, and Resources* (Bonn, Geneva, and Vienna: United Nations University, International Telecommunication Union, and International Solid Waste Association, 2017).
9. Ellen MacArthur Foundation, *Circular Consumer Electronics: An Initial Exploration,* 2018.

08. Smart Energy Infrastructure

Cities, local governments, and communities are trying to manage growth and build for resilience, while much of America's aging energy infrastructure—the electrical grid and gas pipelines, as well as many buildings and transportation systems—needs to be repaired or replaced.

Resilience is a primary concern for the energy infrastructure. It entails the ability to recover rapidly, with minimal interruptions and damage to infrastructure and consumers, when submitted to external stresses such as extreme weather, unexpected outages, or malicious attacks. Such problems dominated the news in the cases of Puerto Rico's power system [1] being crippled by Hurricane Maria in 2017 (the largest blackout in U.S. history) and the destructive 2018 Camp Fire in California (the deadliest U.S. fire in the past 100 years), which was likely started by power lines built in the early 1900s. In 2003, the Northeast blackout [2] left 55 million people in the U.S. and Canada without power for up to 14 days.

Reliable delivery of electricity requires an instantaneous and continuous balance between supply and demand at multiple scales [3]. However, a number of factors are adding increasing uncertainty to the situation, including intermittent renewable energy sources (e.g., solar, wind), more dynamic and unpredictable demand from buildings, increasing use of electric vehicles and evolving charging patterns, and deployment of decentralized power generation/storage facilities [4]. This poses a significant challenge for wide-area coordinated operation of the nation's power grid. These challenges also stem from a lack of flexibility by traditional generation facilities, such as coal-fired and nuclear power plants, to accommodate rapid changes in the supply and demand balance. Moreover, impacts of long-term climate change and short-term extreme weather on the energy infrastructure are intensifying [5]. As the grid continues to evolve at the edge, stationary electrical energy storage has played an important role in the U.S. electricity system. Energy storage solutions currently deployed in the grid, while developed to smooth out peaks or support intra-day shifts in energy consumption patterns, can also be used to integrate electricity from intermittent renewables. Additionally, urban planners have an increasing need for better tools to plan and improve their overburdened transportation infrastructure and co-optimize with electrical infrastructure operators to be ready for future demands, including connected, mixed autonomous, shared, and electrified vehicle fleets.

A smart energy infrastructure that meets energy demands at multi-spatial and temporal scales and operates in an intelligent manner to achieve energy efficiency, flexibility, and resilience is needed to help local, state, and federal governments achieve their energy, economic, and environmental goals. Artificial intelligence can contribute to meeting these objectives.

1. State of the Art

Novel opportunities for AI in this area stem from the rapid deployment of connected devices in the energy infrastructure. Smart energy systems comprise interconnected systems of buildings, urban microclimates, vehicles, power and water supplies, and humans [6]. Urban-scale smart energy infrastructure research offers insights into efficiency, sustainability, and resilience, leveraging emerging opportunities in the Internet of Things (IoT), big data, machine learning, and exascale computing [7]. Modern infrastructure and technologies applied to urban systems include wide-area monitoring, distributed control, advanced communication systems, and varying levels of AI at the edge. IoT has become a critical part of the daily operation in smart buildings, mobility, and the

electric grid, with deployments happening at the city scale. On the transportation side, new data streams from infrastructure sensors and geospatial positioning devices paint a noisy, complex picture of the demands on the transportation system. As a result, large volumes of data are flowing into cities and communities.

Technologies developed for the IoT and smart devices offer an unprecedented opportunity to observe and reliably operate the electrical power system through dynamic control of demand. IoT devices and technologies have effectively provided a “software interface” to energy generation, consumption, monitoring, and control assets that drive the electrical grid, enabling an unprecedented opportunity to federate a large number of heterogeneous devices for performing a decentralized, coordinated control toward a next-generation smart energy infrastructure. Similarly, distributed control of vehicles presents a huge challenge for cities, where currently multiple agencies operate independently with different views of the system and different objectives to optimize.

To design control algorithms that fulfill this potential while accounting for the large numbers of small but now visible and possibly controllable loads, it is necessary to have scalable, data-driven models and understanding that represent the primary elements of generation and transmission, distribution, and the interaction with the primary features of smart buildings. Despite these opportunities, and particularly the data deluge from these devices, AI has been used in a limited way when it comes to energy infrastructure. Typically, various ML techniques were applied to individual buildings or their energy systems, such as virtual sensing (e.g., data-driven models to estimate operational parameters), prediction of thermal and electrical loads, modeling of building energy systems and human-building interactions, detection and diagnosis system operational faults, optimization of control systems, and analyzing human mobility patterns. These efforts are limited to the size and quality of available data, certain energy end uses or single buildings, and single or simplified objective functions. For example, the CityBES [8] tool developed by DOE researchers allows for energy-saving retrofit analysis of hundreds of buildings in a model that considers how the buildings interact (Figure 8.1) [9]. Further, evaluating the impacts of climate change, extreme weather, changing energy usage patterns in buildings, and interactions of transportation and electrical grid are still missing.

Traditionally, resilience (and, consequently, reliability over longer time scales) has been assured by capital- and staffing-intensive activity and massive, repeated offline and online analyses both before and after a major event has occurred. The key “before event” system metrics that can prevent or reduce the severity of an event’s impact have been safety margins, redundancy-by-design capacity (including spare generation and infrastructure), and intense overall situational awareness that includes multiple layers of sensors, external data streams (such as weather forecasts), and state/load estimators. The “after event” metrics have been mostly qualitative indicators of readiness of the local utilities that may be involved in restoration, such as how many training exercises they have participated in and of what type. Since existing data of major event outcomes are exceedingly rare, most of the before-event key metric evaluations and selection are done by means of synthetic data generated by simulation with standard physics and are optimization-based (to emulate the financially driven decision process that takes place outside emergency situations). For system state/load estimation, the models are very crude, typically classical time series models with very simple and often inaccurate models that include a large amount of coarsening and aggregation. The after-event readiness factors are done mostly in an expert-estimated-mode rather than in any form of predictive fashion.

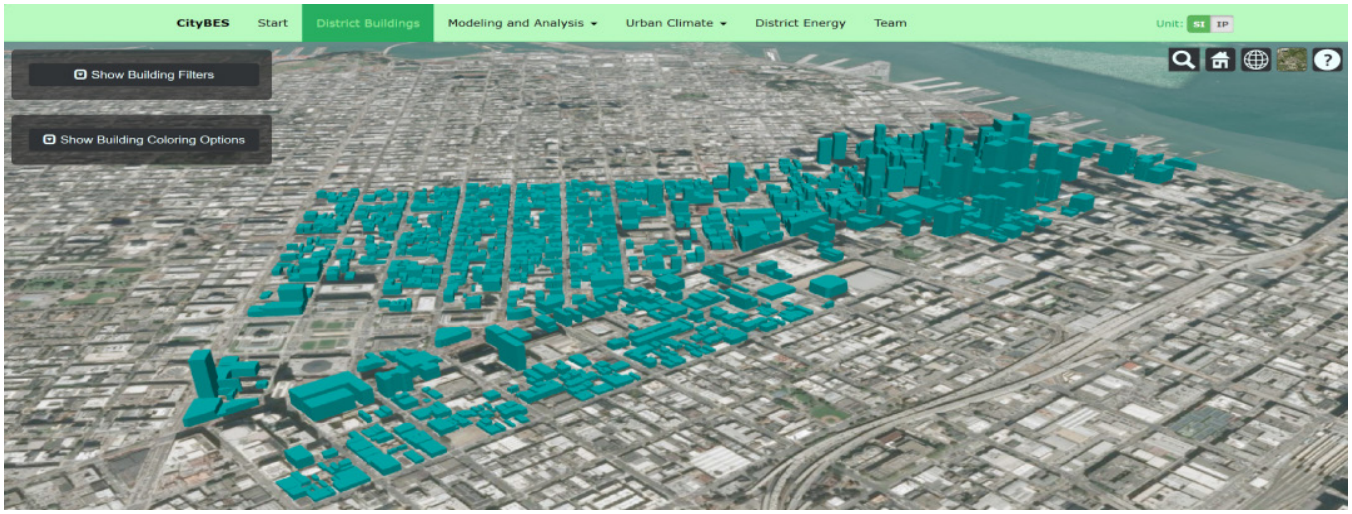


Figure 8.1 *Screenshot from CityBES, a city building energy modeling and analysis tool that considers how the buildings interact and allows for large-scale energy retrofit analysis.*

2. Major (Grand) Challenges

The overarching challenge of the energy sector is providing safe, secure, cost-effective, and clean energy. In the presence of the rapid change drivers of increased variability of supply, reduced inertia, and significantly increased end-user complexity, this requires vastly superior situational awareness and novel tools for rapidly estimating and optimizing the system resilience. Moreover, the multi-stakeholder nature of the energy infrastructure operation, combined with the enormous representation complexity (in the context of distributed urban assets) requires the development of scalable virtualized computational energy infrastructure models (commonly referred to as digital twins) of multiple interdependent energy and consumption infrastructures. These models can serve as support for defining the stakeholder interfaces, help promote optimal usage policies, and enable exploration of scenarios for optimized energy operation and sustainable planning.

Three key grand challenges have emerged for application AI in improving the resilience of the next-generation energy infrastructure as data becomes more pervasive. These grand challenges are:

Wide-area situational awareness to enable energy resilience. Maintaining and increasing the resilience levels in the presence of increased variability of supply, reduced inertia, and novel edge-network structure (such as increased use of microgrids that present the opportunity of increasingly independent operation, but also the challenge of coordination) require much better understanding and prediction of system variability and state. This will enable predictive capabilities of after-event states and sharper awareness during the restoration process. Smart energy infrastructure has the potential to leverage a variety of disparate data sources with varying spatial and temporal resolution to enable AI-driven real-time intelligence for optimal situational awareness. In particular, AI can: a) perform information fusion from disparate data sources coupled with an integrated model of the energy infrastructure at a time scale pertinent to enabling proactive responses to improve resilience, b) enable predictive models for exploring smart energy and transportation infrastructure design, and c) detect and diagnose cyber and physical attacks and threats in real time to ensure security of the energy infrastructure.

Reliable integration of renewable energy into existing infrastructure using distributed optimal control to balance supply and demand. Achieving energy reliability while on the path to clean, sustainable, secure, and affordable energy sources requires high-resolution (spatial and temporal), wide-area distributed control and optimization to balance supply and generation. Simulation and modeling, integrated with AI, provides guidance for wide-area control design. The mechanisms, reliability, and robustness required to deploy control actions for a real-world demonstration will have to be deeply vetted with stakeholders that have operational roles in this sector. A particularly important direction at the intersection of the situational awareness and wide-area control is one of advanced, very high resolution modeling that can represent their switched dynamics without drastic simplification of the generation, delivery, and consumption models. Here AI can offer novel techniques, including surrogate models, closure models, and learning-driven compute acceleration of high-fidelity models and solvers. This research will build on current DOE programs, such as the Grid Modernization Initiative, as well as the Building Technologies Office's research to develop grid interactive efficient building technologies [10], and the Vehicle Technologies Office's Energy Efficient Mobility Systems [11] for affordable and safe mobility.

Fully virtualized urban-scale infrastructure to co-optimize urban mobility and energy end-use. Creating digital replicas of urban mobility infrastructure over a geographic area that is coupled with energy infrastructures allows us to understand, predict, and co-optimize efficiency of energy and mobility infrastructure. A geospatial visualization of deployed sensors feeding in real-time data, as well as mobile sensors that capture trajectories of vehicles and humans, enables the creation of a capability to anticipate future system state and evaluate the impact of control decisions faster than real time. Data sources include signals, sensors, safety information systems (911 and 511), and modern probe data traffic feed from third parties. Starting with the detection of threats/disruptions, the next frontier is to anticipate and mitigate the adverse effects faster than real time of future system states. AI provides solutions to urban-scale challenges, including real-time model training, focus on rare event prediction performance, the multi-stakeholder nature of the data access, and decision rules and procedures that are affected by the complexity of the "what-ifs" that are combinatorial and partially graph-indexed in nature.Opportunities for integration exist at multiple scales, from the drivetrain to vehicle-to-vehicle and vehicle-to-infrastructure exchanges at city and regional scales. A feedback loop from the real world back into the digital replica allows for a level of automated response to perturbations in the network. This requires a deep understanding of how technological disruptions affect human decisions and, in turn, alter demands on mobility.

Developing digital twins of urban systems is a means to address these challenges and provides insights or solutions using AI methods in:

1. Understanding and quantifying the inter-dependencies between buildings, urban climate, transportation, and the grid at multi-spatial (from city block to district to neighborhood to a city to a region) and temporal (from minute to hour to day to month to year to decade) scales under typical or extreme/disrupted situations.
2. Developing strategic pathways to address grid needs beyond the daily cycling and provide backup power for several days that could enhance resiliency by integration of long-duration distributed energy storage systems coupled with renewables.
3. Creating smart operations and controls to integrate buildings and transportation to harmonize with the smart grid for maximal productivity, energy efficiency, demand flexibility, and resilience.

4. Predicting urban systems' dynamic evolution under extreme weather events and understanding how the urban landscape interacts with the microclimate.
5. Detecting patterns of human mobility and charging needs of future autonomous EV.
6. Informing sustainable and resilient urban planning and policymaking, considering long-term population and economic growth as well as climate change and extreme weather events.

3. Advances in the Next Decade

More data (static and dynamic, measured and simulated, physical and human) at the scales of peta- to exabytes from diverse sources will become available and will be integrated into open and interoperable platforms to power the digital twins of smart energy infrastructure. Hardware for edge computing enables migrating the low-level "twin" to the edge for better responsiveness and uninterrupted operations at local devices, which feed information for a higher-level "twin" that implements predictive analytics more efficiently. Federated instrumentation can enable novel softwarization of energy devices to enable scalable information fusion and decentralized control of assets in a reliable fashion. Supercomputing empowered with AI engines will model and simulate smart energy and transportation infrastructure systems as a cyber-physical and natural-human combined system capturing realistic behaviors within the digital twin. Real-time 3D GIS-integrated visualization, coupled with virtual reality and augmented reality in the digital twin, reveals real-time performance of urban systems and pinpoints hotspots (e.g., energy, heat, air pollution, traffic, population, wind). Using it to create a virtual replica of reality would help in building future cities to meet challenges such as extreme weather events and housing and transport needs.

Techniques and knowledge can transfer ML results from data-rich environments to data-poor environments. Each city will have a digital twin that evolves with time, more data, computing power, ML algorithms, and changing environmental and human needs. However, to realize this vision, it is necessary to develop engineering tools, and particularly simulation models and data analytics, that can be used to understand the effects of any particular control strategy and operating circumstance on the now-coupled, parameters of consumption, generation, and power delivery performance. To enable situational awareness and solve resilience-oriented challenges, AI will enable a new paradigm of emerging technology with capabilities to:

- Expand co-simulation tools to include generation, transmission, communication, and increasingly accurate description of the distribution and behind-the-meter areas for enabling fine-grained understanding of system behavior.
- Deliver increased levels of flexibility and control at the lower levels of the system hierarchy, such as the introduction and coupling of microgrids and smart energy management systems with the ability of islanding in the presence of rapidly evolving threats, e.g., fires (see Chapter 15, AI at the Edge).
- Leverage data from rapidly expanding networks of sensors, such as advanced meters, phasor measurement units (PMUs), F-NET, and other sensing and actuation technologies, to revolutionize monitoring and control of power grid (see Chapter 12, Data Life Cycle and Infrastructure).
- Generate HPC-powered simulators of multi-mode mobility of mixed autonomous, electrified vehicles and their interaction with the power grid.
- Integrate multi-physics data sources, with a particular focus on weather measurements and higher resolution weather forecasts (see Chapter 2, Earth and Environmental Sciences).

4. Accelerating Development

The promise of AI addressing these critical challenges can be vastly improved by both institutional and technical accelerated pathways.

Institutional acceleration. This domain features large-scale energy infrastructure and complex systems that will require researchers to establish partnerships with cities, local governments, electric utilities, industry, and others to develop research testbeds based around public and private partnerships. The emerging AI technology needs to demonstrate that it can provide a new understanding of energy infrastructure and provide actionable information for energy system planning, design, and operations. Current demonstrations in this area tend to focus on buildings, transportation, or smart grid technology, individually. Integrating all three of these domains is difficult and will require multi-disciplinary teams.

Piloting fully virtualized, data-driven, computational digital twins of cities is needed to test and validate new technologies and the energy performance of integrated urban systems (see Chapter 15, AI at the Edge). The initial scale could be a city block, then expanding to a district and later to a small city.

One early task would be to evaluate the availability of existing data, data gaps, new sensing and measurement needs, the current state of the art in regional demonstrations, and lessons learned in recent research. It is necessary to determine how AI can and should be coupled with current and future high performance computing simulation capabilities (see Chapter 10, AI Foundations and Open Problems). Examples of the resulting enhanced capabilities are providing more realistic models of complex agent or system behavior inside the models (inner loop) or discovering more optimal control strategies (outer loop) for the system through simulation. In the long run, the research community, in collaboration with stakeholders, needs to demonstrate that it has a vision for how this research can shape the understanding of technology needs, capabilities, priorities, integration opportunities and control, energy performance, and economic value.

This research initiative should provide actionable intelligence to the energy industry to identify gaps in observability to enable deployment of key data sources, platforms for real-time situational awareness and understanding, and novel decentralized control of energy assets in real time.

Technical acceleration. Improved situational awareness across multiple interdependent energy infrastructure requires increased accuracy and resolution of external and derived data streams (heat, mass, urban structures and surfaces, vegetation, weather, and traffic flow). Novel AI algorithms are needed to perform information fusion from disparate data sources coupled with integrated models of the energy infrastructure at a time scale pertinent to enable proactive responses to improve system-wide resilience (see Chapter 10, AI Foundations and Open Problems). The availability of data imposes an increased focus on deployable online learning algorithms, particularly for the restoration process where new data can be extremely informative, to enable multi-scale simulation with AI for model discovery. This domain requires developing novel cooperative AI methods that can support real-time, multi-stakeholder, multi-scale decision-making for the national energy infrastructure. Key technological improvements are needed, with increased focus on human-infrastructure interaction characterization and prediction for transforming largely reactive approaches in use today into proactive resilient operation of the future. Particular emphasis needs to be placed on a sharp characterization of the performance of AI tools for the rare event portion of the prediction space, as seen in recent major natural disasters. This requires development and usage of surrogate models

and understanding emergent behaviors of interacting AI agents that capture the multi-physics of urban systems and can learn from the combination of measured data and physics- or model-based simulation data for rapid prediction. Real-time forecasting techniques have to be developed by leveraging urban sensing and monitoring (e.g., building energy use and system operation, traffic flow) to predict operational issues (e.g., traffic, power demand) and inform preventive actions. Key AI-driven optimization methods that are applicable for control deployment to operate in real time with deep reinforcement learning have to be developed to achieve optimal performance of the complex urban systems delivering multiple objectives (efficiency, flexibility, and resilience) under uncertain real-world conditions.

5. Expected Outcomes

AI will enable the development of high-resolution situational awareness and resilience-focused control of smart energy infrastructure by combining diverse data sources and creating novel models and synthetic datasets spanning multidisciplinary sciences (building science, urban science, mobility science, sensing and communication, data science, AI, computing science, behavioral/decision science). The models and datasets will deliver data-driven decision support to address grand challenges of urban energy and environment, considering interconnected systems of buildings, climate, transportation, smart grid, and humans. They will also support real-time operations and optimization of integrated urban systems by means of computationally efficient optimization through intelligent interacting AI agents. The fully virtualized data-driven models of smart energy infrastructure will enable stakeholders, decision makers, and citizens to benefit from efficient, flexible, and resilient operations under normal, stressed, and extreme conditions. We believe that AI can improve the resilience significantly—possibly by an order of magnitude compared to a business-as-usual approach.

6. References

1. Kwasinski, F., Andrade, M. J. Castro-Sitiriche and E. O'Neill-Carrillo, "Hurricane Maria Effects on Puerto Rico Electric Power Infrastructure," *IEEE Power and Energy Technology Systems Journal*, **6**, 85-94 (2019). doi: 10.1109/JPETS.2019.2900293
2. U.S.-Canada Power System Outage Task Force, *Final Report on the August 14, 2003 Blackout in the United States and Canada: Causes and Recommendations*, April 2004.
3. IEEE guide for electric power distribution reliability indices, IEEE Std 1366-2012.
4. U.S. Department of Energy website, "Confronting the Duck Curve: How to Address Over-Generation of Solar Energy," (2017).
5. National Academies of Sciences, Engineering, and Medicine 2017. Enhancing the Resilience of the Nation's Electricity System. Washington, DC: The National Academies Press. doi.org/10.17226/24836.
6. Hong, T., Chen, Y., Lee, S.H., Piette, M.A. "CityBES: A Web-based Platform to Support City-Scale Building Energy Efficiency," *Urban Computing* (2016).
7. Chen, Y., Hong, T., Piette, M.A. "Automatic Generation and Simulation of Urban Building Energy Models Based on City Datasets for City-Scale Building Retrofit Analysis," *Applied Energy* (2017).
8. U.S. Department of Energy, "Smart Grid System Report," (2018).
9. Hong, T., et al. "Ten questions on urban building energy modeling," *Building and Environment* (2019).
10. U.S. Department of Energy, Grid Interactive Efficient Buildings https://www.energy.gov/eere/buildings/grid-interactive-efficient-buildings
11. U.S. Department of Energy, Energy Efficient Mobility Systems https://www.energy.gov/eere/vehicles/energy-efficient-mobility-systems

This page intentionally blank.

09. AI for Computer Science

Artificial intelligence methods were originally developed to solve one of the grand challenges in computer science, namely the design of computer systems that could behave like humans. The most recent breakthroughs in AI use machine learning to address specific problems in computer vision, natural language processing, and robotics, and to outperform human players in games of strategy like chess and Go. AI has the potential to address a variety of computer science challenges where complex manual processes could be replaced by automation, including chip design, software development, and online monitoring, and decision making in operating and runtime systems, database management, and infrastructure management.

The DOE Office of Science Advanced Scientific Computing Research (ASCR) program drives innovations and improvements in scientific understanding through its world-class research program and facilities—both computing and networking. The innovations in science user facilities (see Chapter 14, AI for Imaging) are expanding the boundaries of computing to include the edge (see Chapter 15, AI at the Edge), consisting of science instruments and sensor networks (see Chapter 16, Facilities Integration and AI Ecosystem). Traditional computer science will not be sufficient to address the complexity and scale of future systems and workloads arising in the DOE science mission described in Chapters 1 through 8. AI will provide solutions to the design, development, deployment, operation, and optimization of all hardware (see Chapter 13, Hardware Architectures) and software components (see Chapter 11, Software Environments and Software Research), ranging from individual elements to coordinated orchestration of the workflows over computing, networking, and experimental facilities.

In this chapter, we identify the grand challenges in computer science that can be addressed by AI. Specifically, we identify grand challenges in the areas of hardware and software system design, programming, theoretical computer science, and workflow and infrastructure automation. We do not address computer science solutions to support AI, which is covered in other chapters (see Chapters 11–13).

1. State of the Art

AI has the potential to transform many fields of computer science, from low-level hardware design to high-level programming and from the most fundamental algorithmic challenges to day-to-day operation of user facilities.

Hardware and software design. The design of next-generation hardware and software systems and mapping of application codes to target systems is currently a static process that involves human-in-the-loop design processes and consists of repeated experiments, modeling, and design space exploration. The design of new chips and HPC systems takes many years, and hardware vendors and application developers spend months mapping, porting, and tuning applications to run on new systems. As hardware and software get more complex and heterogeneous, current strategies will be impractical. DOE has been a leader in the co-design of HPC systems for science, but many hardware features are still driven by technology constraints and can be a challenge for programmers (see Chapter 11, Software Environments and Software Research). The DOE community has also spearheaded the use of automatic performance tuning (autotuning) using both brute force search and mathematical optimization [5–8]. In recent years, AI has been explored for the design of chips [1], storage management [2], hardware [3], optimizing compilers [6,7], and to improve the performance of single-node computation [5,8], communication, I/O [9,10], math libraries [15], and scheduling [11]. However, the payoffs

from AI-driven hardware and software co-design are far from complete and will require rapid and non-intrusive data collection, exploration and development of methods, and sharing of learned models.

Application development and data wrangling. Development, tuning, maintenance, and testing of software and making data ready for models and methods are manual, expensive, tedious, and error- prone processes (see Chapter 11, Software Environments and Software Research and Chapter 12, Data Life Cycle and Infrastructure). Existing techniques for developing software were mostly designed for logic-heavy control flow programs that run on a single machine with homogenous hardware. However, the future of software demands data-driven, distributed programs that run efficiently on heterogeneous hardware. Recent work in program synthesis and automated testing has produced tools that work independently to generate and test software or as powerful aids for human developers (Figure 9.1).

Automated program synthesis produces software solutions based on either input/output examples, demonstrations, or high-level specifications, producing software from equations, code written in a domain-specific language, or a simple unoptimized version of a program [4,12,14]. Programmers employ various approaches to tackle complex coding tasks, including Google search and online communities, and use integrated development environments that help them to autocomplete code, which can be partially automated with natural language code search [19] and code recommendation [18]. Smart fuzzing techniques, which use random or invalid input to test a computer program, have shown promising results in helping to find semantic bugs in large software systems. Programming by optimization [16] is a design paradigm that allows software developers to specify a rich and potentially large design space of software components that can be used by AI to generate programs that perform well in a given context.

Data wrangling today is largely a human-intensive task, and AI offers unique opportunities to automate or simplify the task. For example, ActiveClean [22] provides a set of optimizations to select the best data to be cleaned for an iterative cleaning framework.

Computer science foundations of AI. AI methods have been increasingly applied to solve complex science problems using codes

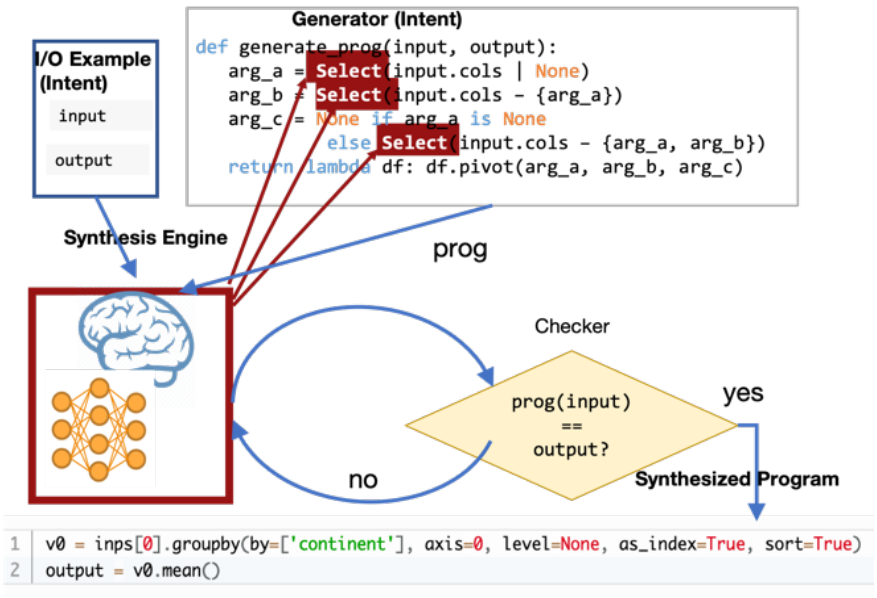


Figure 9.1 *AutoPandas uses neural-backed operators in program generators for program synthesis [4].*

with provable performance and correctness properties. It has been shown that certain smoothness and boundedness properties of physical and abstract system laws can be exploited to develop domain-specific, effective ML solutions with many desirable properties (see Chapter 10, AI Foundations and Open Problems). For example, they lead to performance optimization of data transport infrastructures [13] and accurate power-level estimation of reactors [28]. Principles of theoretical computer science provide a rigorous framework to establish critical properties of AI/ML codes, namely computability, learnability, explainability and provability, as illustrated in Figure 9.2.

There are several known performance limits of ML methods, and many practical problems have shown to be within them. Indeed, several critical problems—such as zero-day computer virus detection [26] or assessment of code resilience to arbitrary hardware faults [24]—are non-Turing computable and hence not solvable by black-box ML methods. In some cases, the complexity of the tasks could be too great—such as the unbounded Vapnik-Chevenenkis dimension [25]—so that no performance guarantees can be given for any ML solution, independent of the sophistication of its design or use of a supercomputer. It appears on the surface that limitations of "black-box" ML solutions can be overcome by requiring that they be explainable, but Tarski's limit prevents a machine from generating explanations in some cases even if they exist [23]. More recent results show that learnability may be undecidable [29], and similar results are expected to appear in the future that establish limits of ML and its performances. Complementing more general considerations in Chapter 10, the theoretical computer science provides the frameworks and tools to establish that a given problem indeed is effectively solvable by AI/ML methods and is not subject to the above limits.

Workflow and infrastructure management. Managing distributed infrastructure that spans multiple systems, domains, and organizations is largely achieved today by manual or ad-hoc methods for configuration, monitoring, and optimization (see Chapter 14, AI for Imaging and Chapter 16, Facilities Integration and AI Ecosystem). Challenges exist at multiple levels, from assuring the safe and secure operation of networks and systems to efficient resource allocation to users and optimal use of the distributed systems for complex scientific workflows. Neither individual users nor system operators have the global view or integrated control mechanisms needed to make efficient use of a multi-purpose, multi-facility infrastructure. AI provides the automation that can ease the burden of human-driven management of infrastructure at facilities.

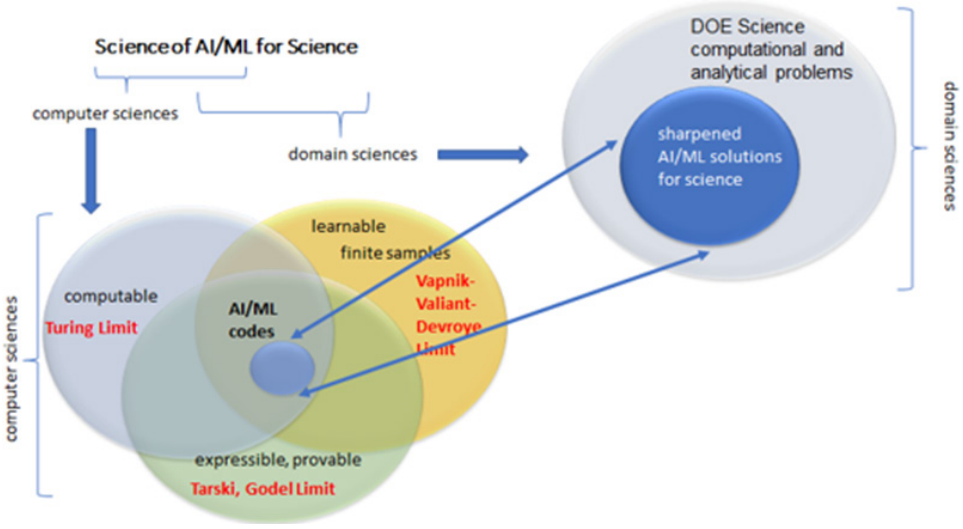


Figure 9.2 *Computable, learnable, explainable, and provable AI/ML solutions.*

Current efforts are exploring the use of reinforcement learning, unsupervised learning, and classification techniques to optimally control wide area network resources (Figure 9.3) to improve high-speed big data transfers and analyze the performance variation of applications on supercomputers to correlate with key users and workloads and leverage software-defined networking and related services exported in high-level programming interfaces, incorporating them into complex workflows [13]. The future will be automated, policy-driven management of distributed resources by both individual applications and complex-wide workflows.

2. Major (Grand) Challenges

We identify three grand challenges in the areas of hardware and software system design, programming, workflow and infra-structure management, and foundational computer science. The grand challenges together attempt to optimize productivity, performance, reliability, and portability across the DOE complex.

Develop hardware/software systems that are semi-automatically co-designed and co-tuned. In recent years, heterogeneous hardware and software infrastructures have been deployed at DOE high performance computing facilities. Additionally, a number of different accelerators are emerging in the community, including neuromorphic and quantum computers (also see Chapter 11, Software Environments and Software Research). However, the cost and time required for the design of current HPC hardware and software is still prohibitive. AI can influence the co-design of hardware and software systems at many levels to meet energy, security, resilience, and performance requirements. For example, at the transistor level, it can find an optimal set of device parameters given a set of operating constraints; at the chip level it can optimize placement of logic blocks and configure architectural-level features such as the number of floating-point units. AI can be used at runtime to monitor hardware characteristics such as thermal densities to anticipate faults that deliver incorrect results or disable parts of chips. It can also be used in compilers, programming libraries, and applications to automatically generate and search through various implementations to find one optimized for a given hardware platform, and even adapt dynamically to changing performance characteristics at runtime. For example, AI could identify common parallelization of memory optimization strategies and transfer them across applications.

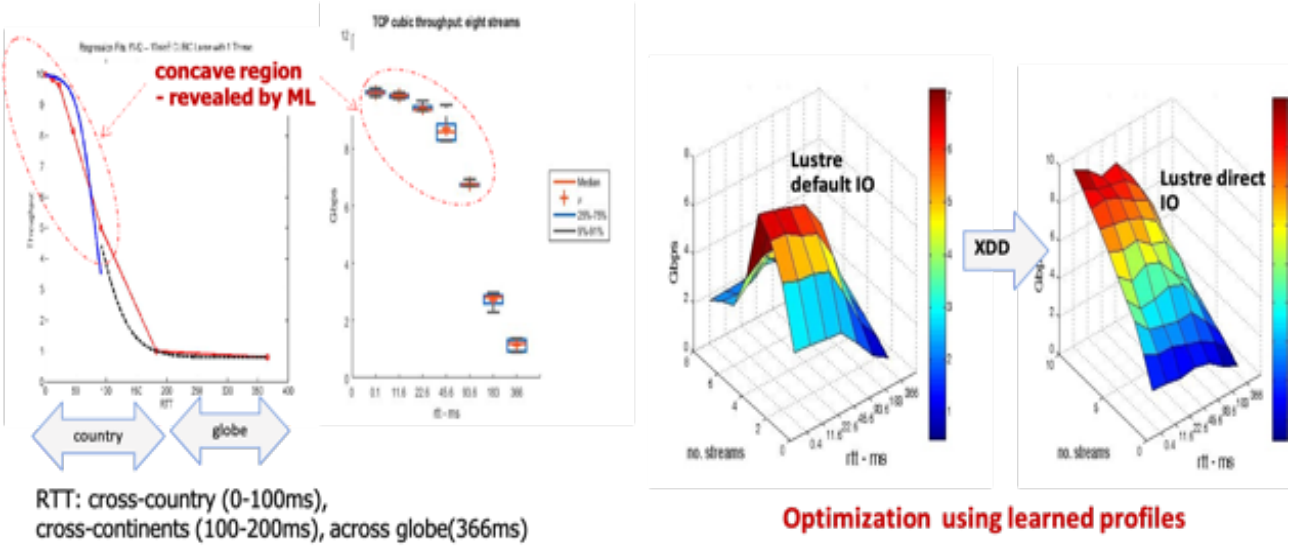


Figure 9.3 *The performance of data transfer infrastructure depends on all its subsystems, namely, networks, transfer hosts, file and I/O systems, storage systems, and data transfer. Custom ML methods have been developed to estimate throughput profiles [13].*

AI methods can be leveraged in operating systems and runtime systems. They can monitor applications running on large-scale HPC systems and learn a model of application performance. These models can provide feedback on how to best manually or automatically adapt the application for better performance. AI-based performance modeling can capture data flow and dynamic performance behaviors of parallel applications under various constraints; model-based analysis will extract concurrency, examine the tradeoffs among different performance factors, and make predictions about different applications on future systems. Model-based optimization will allow users to optimize the performance of applications based on target objectives. In addition, AI methods can be applied to assist model construction, facilitate performance analysis, accelerate optimization, and provide software verification.

Enable automated tools for programming and data wrangling suited for modeling and data-driven science needs. Automated and computer-aided programming tools will be developed by using AI-driven program synthesis and code recommendation, software adaptation, software testing and verification, and code optimization. This will dramatically reduce or eliminate programming efforts for high-level applications on heterogeneous architectures and will support a new generation of programmers for data analysis and learning, demonstrating that HPC novices can accomplish tasks in a tenth of the time that experts spend with traditional tools, produce programs that are 10 times faster than expert-written programs, and increase automated test coverage to 95% or more.

In the future, programmers will perform complex programming tasks by expressing their intents via high-level domain-specific languages, input-output examples, demonstrations, natural language descriptions, and formal specifications. Program synthesizers will then take such intents and search combinatorically large spaces of possible candidate programs. AI-driven program synthesis will learn heuristics by extracting probability distribution of programs from real-world corpus of programs and by remembering search strategies that worked well in the past. AI-driven code recommendation, such as auto code completion, will help find the right libraries and APIs and synthesize or recommend new code using these libraries and APIs. Additionally, automated techniques will extract intents from user inputs and adapt them to different environments. Automated testing based on smart fuzzing code perturbations and dynamic symbolic execution will allow developers to efficiently test code.

Enable automated and efficient execution of end-to-end scientific workflows processing experimental, observational, and simulation data on adaptive and resilient infrastructure. We envision a future in which a researcher at a user facility would be able to launch his or her experiment and seamlessly access the network and resources in real time at HPC facilities to process data, compare with simulation results, search other relevant data, and reproduce the workflow. Future novel workflows may include AI elements combined with simulation and experimental science.

A recent ASCR workshop report lays out the challenges and approaches for using ML to develop distributed, fault-tolerant, energy-efficient HPC applications [17]. An AI-driven autonomous workflow engine will use user input and prior learned knowledge of the system to generate optimized code, use the workflow through intelligent schedulers that use AI in addition to policies, and monitor the execution. AI can guide scientists in designing and optimizing their workflows in ways that are not possible today. These workflows will run atop a fully automated infrastructure in which AI will design, develop, deploy, monitor, diagnose, operate, and optimize computing elements, units, systems, complexes, networks, databases, and federations. The end user at the user facility and the facility staff at ASCR facilities will be notified of situations that

require visualization or, more generally, humans in the loop. The autonomous workflow and all data associated with it will also be captured and made available to be published in machine-readable journals. AI provides a unique opportunity to automate the management of the underlying infrastructure and the scientific process to accelerate the pace of scientific discoveries.

Develop computable, provable, explainable, performance-guaranteed and yet practical AI solutions for science. AI/ML methods that exploit the properties and structure of underlying system and abstract laws lead to customized solutions that are computable, explainable, and possess proven generalization and correctness. In particular for science problems, such solutions range from inferring new inter-relationships, efficient polynomial approximations to NP-hard problems, discovering new laws from measurements and simulations, and obtaining optimizing parameters over complex spaces. There is an immediate need for foundational frameworks and tools that enable us to assert the critical properties of AI/ML solutions by combining the rigorous theories of computing, learnability, expressability, inference, and provability, which have been developed as highly specialized individual technical areas. They have to be refined, sharpened, and combined to address the spectrum of science areas consisting of interacting physical and cyber systems, such as simulation-driven experiments, experiment-steered computations, and optimal design and operation of smart grids and federations of computing systems and experimental facilities. The underlying laws here are hybrid in encompassing both systems, physical and cyber. Establishing that the solution is indeed within ML foundational limits, and exploiting the properties of underlying systems across the myriad of DOE computational tasks, are challenges considering the diversity of science areas in which ML methods are being applied.

3. Advances in the Next Decade

In the short term, AI can be an invaluable tool for analyzing observational data in computing systems, applications, facilities, and networks. Over the next decade, AI techniques will detect and anticipate performance anomalies due to hardware failures, resource overload, intrusion, or other interactions. It will accelerate the design of hardware and software through intelligent design space exploration and improve the automated tuning of high-performance libraries and applications. Longer term, the analysis will be used for online learning and real-time control of increasingly sophisticated application workflows that cohesively tie together the facilities and other resources across DOE and the broader science complex.

The grand challenges rely on innovations at different spatial scales, including the node, machine, and facility level. At the node level, the impact of various hardware and software knobs will produce learned multi-metric performance models for runtime, power, energy, memory footprint, and more. At the machine level, AI will learn models of communication, load balancing, and I/O and try to understand the impact of resource sharing across applications. Facility-level models will capture resource utilization, power constraints, user satisfaction, and time-to-solution to optimize job scheduling, staging phases of the application, and file transfers. Model-informed decisions will be made at every level with multiple coordinated feedback loops. In particular, the coordination will happen both bottom up (node to facility) and top down (facility to node).

Reliance on ML in DOE science and energy applications, user facilities, and cyber-physical systems means there is a new part of the system that can be attacked, such as via tainted training data, false sensor data, and fragile AI algorithms. Any use of AI, particularly

AI-automated processes, is vulnerable to such attacks. Consequently, detecting tainted training data and false sensor data, and measuring confidence of AI algorithms in their output become critical as these AI-enabled systems are deployed. While the existing methods are primarily ad hoc and heuristic in nature, recent methods include the development of AI-based cybersecurity methods. For example, adversarial training [20] is an approach that injects adversarial examples into training data to increase robustness of ML models. For the Cybersecurity of Cyber-Physical Systems and DOE facilities, the conventional methods have become inadequate (for example, zero day threats), and new AI-based cybersecurity mechanisms are under active development [21].

4. Accelerating Development

Many ongoing efforts are using AI to address challenges in computer science. However, strategic investments and coordinated efforts at both technical and programmatic levels will be needed to realize the vision outlined in these grand challenges.

Access to curated data from many different levels of hardware and software is key. Data analogous to ImageNet is needed to feed AI for computer science. This is beyond the capacity of the individual researcher and must capture various design aspects of hardware, software, programming, workflow, and infrastructure. There are a number of challenges to collecting this data, including applicability (i.e., using data from an old system for a new system might not result in meaningful predictions), and accessibility (i.e., data needs to be extracted and available in formats that are meaningful to the models). ASCR facilities already have organized efforts to make available more data and will pave the way for more autonomous infrastructure. AI for programming today can benefit from websites such as GitHub and Stack Overflow that offer massive amounts of data and metadata about programs. Efforts will be needed to identify data and software repositories that are specifically applicable for the scientific community.

We need to develop open-source scalable modeling and simulation for the entire infrastructure to test AI algorithms. We need methods and algorithms that can operate at different scales; for instruction scheduling to pointer chasing, we need lightweight, low-latency ML methods, but for system co-design we need ML methods that can scale to full exa/zetta scales to explore the vast design space parameters. Additionally, AI will need appropriate infrastructure. Cloud computing platforms have been a cornerstone for scaling ML/DL and AI methods in industry. Similarly, the use of HPC and other platforms to support AI workloads will be critical (see Chapter 13, Hardware Architectures).

Enabling autonomous workflows on autonomous infrastructure will take years of sustained efforts across experimental, computing, and networking facilities. The realization of this vision will require a fast, dynamic, optimized, distributed software-defined ecosystem, critical measurement streams, and powerful analytics that can extract information to drive allocations, diagnosis, and broad strategies and policies. Initial efforts can focus on automated, adaptive collection of instrumented data from many devices and at multiple levels as well as AI-driven integration into dynamic composite state. Automating parts of the workflow (e.g., resource allocation) based on historical data will enable us to lay the foundation for autonomous workflows. Additionally, sustained performance optimization using trend detection, strategy adaptation, continuous performance monitoring, predictive diagnosis, and graceful task reallocation and migration using AI methods will provide starting points for this work.

Programmatically, we suggest a number of efforts to realize the grand challenges. A hardware-software co-design effort that includes researchers, industry partners, and

the computer facilities will be needed to achieve the grand challenge of developing self-improving and self-adaptive hardware-software systems that can be designed and operated without significant human involvement. Teams of computer science researchers working closely with experimental, computational, and networking facilities to develop autonomous workflows on autonomous infrastructure will be needed.

The foundational computer science challenge will require a comprehensive AI/ML science program (across math and computing science) to develop and refine foundational limits and solvable problems and to sharpen the solutions for solvable classes to ensure effective computation, performance guarantees and explanations. The program would benefit from a SciDAC-style consortium for domain scientists working closely with ML scientists to act as a DOE-wide, central resource to be used to analyze the ML problems, establish their solvability, and develop effective solutions. Finally, it will be critical to retrain existing staff and hire and retain new talent with expertise in various areas of computer science, including distributed infrastructure, AI, and foundational computer science.

5. Expected Outcomes

The use of AI will allow us to address hard challenges in computer science toward automating human-intensive parts and reducing time to innovations in hardware, software, workflows, and infrastructure to meet the utilization and performance needs while enhancing scientific productivity. These innovations will directly impact scientific discovery, allowing users to set up federations and execute workflows on well-oiled infrastructures. AI will directly result in optimal facility utilization and response—self-healing, self-optimizing infrastructures will handle the predictable problems while human operators will have the tools to diagnose and fix problems.

6. References

1. Ibrahim, A., Elfadel, M., Boning, D., Li, X. (Ed.), Machine Learning in VLSI Computer-Aided Design, Springer International Publishing, 2018.
2. Toigo, J., AI for Storage Management Gets Real. *Tech Target* (2019). https://www.google.com/amp/s/searchstorage.techtarget.com/opinion/AI-for-storage-management-gets-real%3famp=1
3. 2nd International Workshop on AI-assisted Design for Architecture https://eecs.oregonstate.edu/aidarc/index.php/program/
4. Bavishi, R., Lemieux, C., Fox, R., Sen, K., & Stoica, I., AutoPandas: Neural-Backed Generators for Program Synthesis. *Proceedings of the ACM on Programming Languages*, OOPSLA'19, October 2019.
5. Ansel, J., Kamil, et al., Opentuner: An extensible framework for program autotuning, *Proceedings of the 23rd International Conference on Parallel Architectures and Compilation*, 303–316. ACM, 2014.
6. Balaprakash, P., et al., Autotuning in High-performance Computing Applications, *Proceedings of the IEEE*, 1–16, 2018.
7. Tiwari, A., Chen, C., Chame, J., Hall, M., & Hollingsworth, J., A Scalable Auto-tuning Framework for Compiler Optimization, *Proceedings of the 2009 IEEE International Symposium on Parallel & Distributed Processing*, 1-12, 2009.
8. Thiagarajan, J. J., et al., Bootstrapping Parameter Space Exploration for Fast Tuning, *Proceedings of the 2018 International Conference on Supercomputing*, 385–395, November 2018.
9. Marathe, A., et al. Performance Modeling Under Resource Constraints Using Deep Transfer Learning, *Proceedings of the International Conference for High*

Performance Computing, Networking, Storage and Analysis (SC17), 31, 2017.

10. Behzad, B., et al., Taming Parallel I/O Complexity with Auto-tuning, *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis* (SC13), 68, 2013.
11. Lin, X., Wang, Y., & Pedram, M., A Reinforcement Learning-based Power Management Framework for Green Computing Data Centers, 2016 IEEE International Conference on Cloud Engineering (IC2E), 2016.
12. Kalyan, A., et al., Neural-guided Deductive Search for Real-time Program Synthesis from Examples, The Sixth International Conference on Learning Representations (ICLR 2018), 2018.
13. Rao, N. S. V., Sen, S., Liu, Z., Kettimuthu, R., & Foster, I., Learning Concave-convex Profiles of Data Transport Over Dedicated Connections, Machine Learning for Networking, Springer-Verlag, 2019.
14. Cai, J, et al., Making Neural Programming Architectures Generalize Via Recursion, The Fifth International Conference on Learning Representations (ICLR 2017), 2017.
15. Sid-Lakhdar, W., Mahmoudi Aznaveh, Mohsen, M.A., Li, X., & Demmel, J., Multitask and Transfer Learning for Autotuning Exascale Applications, submitted August 2019.
16. Hoos, H.H., Programming by Optimization. *Communications of the ACM* **55**, 70–80 (2012). DOI: https://doi.org/10.1145/2076450.2076469
17. Berry, M., et al., *Machine Learning and Understanding for Intelligent Extreme Scale Scientific Computing and Discovery*, technical report, DOE ASCR Workshop Report, 2015.
18. Luan, S., Yang, D., Barnaby, C., Sen, K., & Chandra, S., Aroma: Code Recommendation via Structural Code Search,*, Proceedings of the ACM on Programming Languages* (OOPSLA'19), October 2019.
19. Cambronero, J., Li, H., Kim, S., Sen, K., & Chandra, S., When Deep Learning Met Code Search*, Industry Track of 27th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (ESEC/FSE'19), ACM, 964–974, August 2019.
20. Tramèr, F., et al. Ensemble adversarial training: Attacks and defenses. arXiv preprint arXiv:1705.07204 (2017).
21. ASCR Cybersecurity for Scientific Computing Integrity - Research Pathways and Ideas Workshop https://escholarship.org/content/qt5j00n7h2/qt5j00n7h2.pdf
22. Krishnan, S., J. Wang, E. Wu, M. J. Franklin, and K. Goldberg. 2016. ActiveClean: Interactive Data Cleaning for Statistical Modeling. Proc. VLDB Endow. **9**, 948–959 (2016).
23. Tarski, A.. *Logic, Semantics, Metamathematics: Papers from 1923 to 1938*, Oxford University Press, 1956.
24. Rao, N. S. V. On undecidability aspects of resilient computations and implications to Exascale, Resilience 2014: Seventh Workshop on Resiliency in High Performance Computing with Clouds, Grids, and Clusters, 2014.
25. Vapnik, V. N. *Statistical Learning Theory*. John-Wiley and Sons, New York, New York, 1998.
26. Cohen, F. B. "Computational aspects of computer virus," *Computer & Security*, **8**, 325–344, 1989.
27. Rao, N. S. V., Reister, D. B., Barhen, J. Information Fusion Methods Based on Physical Laws*, IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**, 66–77 (2005).

28. Rao, N. S. V., et al. Multi-Modal Sensor Fusion for Reactor Power-Level Estimation: Thermal, EM, Acoustic. Nuclear Security Applications Research & Development Program Review Meeting, 2019.

29. Ben-David, S., Hrubes, P., Moran, S., Shpilka A., and Yehudayoff, A. *Learnability Can Be Undecidable*, Nature, 2019.

10. AI Foundations and Open Problems

Advancing the mathematical, statistical, and information-theoretic foundations of artificial intelligence is vital to realizing the potential of AI for science. These foundations are now a bottleneck for scientific discovery, and the practical application of AI and machine learning remains predominantly an art. Although significant progress is being made, advances in the foundations of AI will be required to complement capabilities in hardware and software and realize the full potential of AI in DOE's science and engineering mission (see Chapters 1 through 9).

One of the distinguishing characteristics of science is the existence of laws based on time-tested observations about natural phenomena. How should these governing principles and other scientific domain knowledge be incorporated in an AI era? To become an accepted part of the toolbox of scientists and engineers, the validity and robustness of AI techniques need to be trusted. What are the limits of AI techniques, and what assumptions and circumstances can lead to establishing assurance of AI predictions and decisions? Another hallmark of science and engineering is that limited training data may be available in the most complex, dynamic, and high consequence of applications. Which AI techniques can best address different sampling scenarios and enable efficient AI on various computing and sensing environments?

Addressing these and other open problems will advance the building blocks of the entire AI ecosystem.

1. State of the Art

Advances in algorithms and hardware have given scientists the tools to model and simulate nature at an unprecedented range of scales: from computing the history and fate of the cosmos and the explosion of supernovae to the evolution of the climate system and the properties of materials to the smallest of subatomic particles. These efforts have traditionally relied on mathematical, modeling, and computational building blocks whose properties are well established. Despite having access to tremendous computational resources, the fact remains that scientists cannot possibly explore all possible theories or simulate phenomena at the sub-grid scale. AI presents a unique opportunity for bridging this gap, but its building blocks and their composition are not yet sufficiently established for widespread scientific use [2–4].

Although the past decade has seen significant algorithmic and theoretical progress, work on the foundations of AI and ML has been far outpaced by the empirical exploration and use of these techniques [16]. With the increased use of AI and ML, clear trends are emerging. For example, residual network-based convolutional neural networks [10,11] are the standard for image processing; automatic differentiation and accelerated first-order optimization algorithms are pervasive in training deep networks [12–15]; and generative models (e.g., generative adversarial networks, variational autoencoders) are providing synthetic data far beyond traditional image applications [6–9,17,25]. Principles underlying the use and understanding of these and other techniques tend to be scattered across disciplines, from theoretical computer science to signal processing to statistics.

Neural networks have started to be specially designed to incorporate some types of domain knowledge—such as rotational equivariance [1,5,21] (Figure 10.1) and statistical [18], partial differential equation (PDE), [19] and stochastic PDE [20] constraints—but these efforts are in their infancy. Results are also being established in the computability of AI-related problems [26] and in exploiting graph-based representations [22–24]. Natural language processing and unsupervised learning

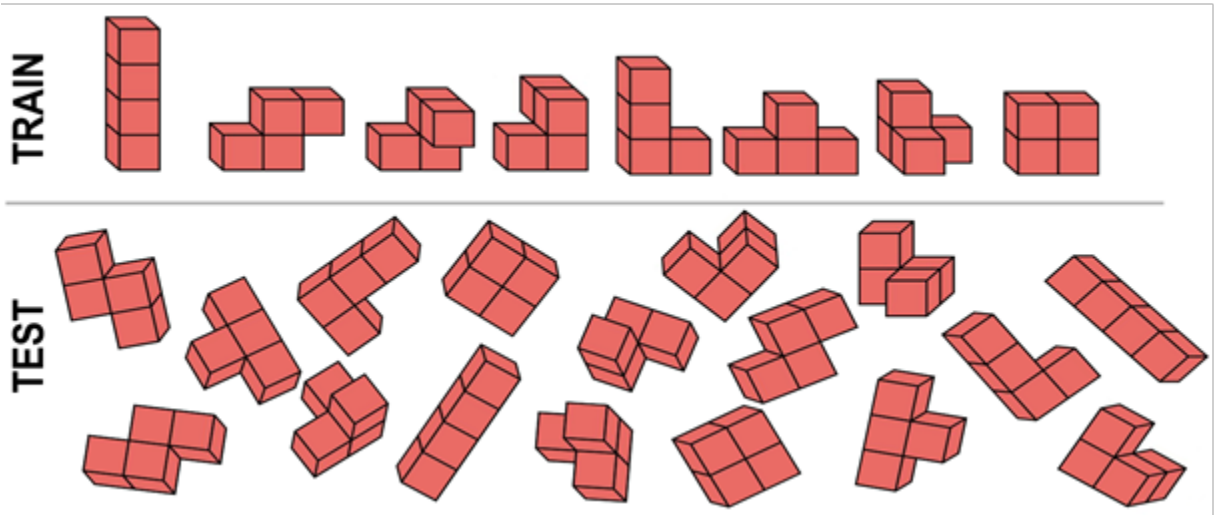


Figure 10.1 *Specially designed neural networks can satisfy domain properties such as 3D rotation-equivariance, allowing one to train on shapes and molecules in one orientation while still identifying shapes and molecules in any orientation. Adapted from N. Thomas, NeurIPS18 Molecules and Materials Workshop [1].*

techniques are beginning to be explored to gain additional insight from the scientific literature [27,28] and to pass eighth grade science exams [29].

2. Major (Grand) Challenges

Three exemplar grand challenges are identified to illustrate the promise of addressing the foundations of AI.

Incorporate domain knowledge in ML and AI. ML and AI are generally domain-agnostic. Whether studying datasets from a beamline scattering experiment, a physics collision, or a climate simulation, the training procedure typically treats every labeled dataset as a point in a high-dimensional space and proceeds to apply standard convolutional and nonlinear operations. Off-the-shelf practice treats each of these datasets in the same way and ignores domain knowledge that extends far beyond the raw data itself—such as physical laws, available forward simulations, and established invariances and symmetries—that is readily available for many systems, much in the same way that early knowledge on the neural vision system led to marked improvements in image processing. Better incorporation and entirely new methods targeting these principles will improve data efficiency; quality, interpretability, and validity of the model; and generalization, transfer learning, and constraint satisfaction for new problem regimes. Incorporating modeling and simulation capabilities to generate training data leverages decades of HPC improvements to accelerate learning; incorporating mathematical equations and scientific literature leverages centuries of advances in theory. Furthermore, to complete the scientific process, incorporating domain knowledge in AI models can be used as the basis for advances in experimental design, active learning, facilities operations, formal verification, and automated theorem proving to accelerate scientific discovery.

Improving our ability to systematically incorporate diverse forms of domain knowledge can impact every aspect of AI, from selection of decision variables and architecture design to training data requirements, uncertainty quantification, and design optimization. Indeed, incorporating domain knowledge is a distinguishing feature of AI within the DOE mission, without which AI-based scientific progress is otherwise limited to that afforded by traditional AI drivers.

Establish assurance for AI. Assurance addresses the question of whether an AI model has been constructed, trained, and deployed so that it is appropriate for its intended use,

and is one of the most challenging problems facing AI. Briefly, it addresses the question of whether and when an AI model can be trusted. Assurance is an extremely broad topic and includes the validity, robustness, reproducibility, and uncertainty quantification of both learned models and their use, as well as the topics of explainability and interpretability. It also includes the question of whether the data used in training an AI model contains sufficient information to train the model without introducing spurious correlations or bias that will invalidate AI-based decisions, as well as operational assurances in the presence of limited/noisy data or adversarial attacks. Furthermore, it includes the development of provable methods to assess whether a problem is computable, learnable, and expressible given the available data and other limitations.

As an example, establishing assurance for a particular AI model would involve clarifying and answering questions such as: Why does the AI model work for a problem? What are the internal representations of data that the AI model has learned during training? How can the behavior of the AI model be explained? How confident are the AI models on their predictions given the different sources of uncertainties and inductive biases involved? For such an AI model to be accepted as a well-characterized tool for science, the research community will need to address these questions and develop advanced capabilities to explain the behavior of the AI model form and map the internal representation of the model to domain-specific concepts.

It would then follow that establishing assurance means determining whether an AI model is appropriately trained and used for the task for which it is intended, including whether it is robust against adversarial attacks or whether prediction errors can be meaningfully bounded. Explainability can also be used to provide the basis of trust in AI systems by communicating meaningful information to humans and for *a posteriori* in diagnosing AI behaviors. Unique challenges for AI systems revolve around a careful characterization of the generalization limits, proofs of validity, robustness, and assessment of confidence associated with predictions. Establishing assurance is especially vital for scientific and high-consequence applications where AI models and tasks would otherwise fail to be adopted, including autonomous systems such as in advanced manufacturing, energy generation, storage and distribution, automated health diagnostics, and the control of large scientific facilities.

Achieve efficient learning for AI systems. The core of any AI system is the creation of an abstract model and the training of that model based on data. Efficient learning in ML systems must be studied along several axes. The first is algorithmic. For example, deep neural networks routinely include hundreds of layers and billions of trainable parameters. Training these models for complex applications is computationally intensive, requiring large amounts of computing power and data, and is critically dependent on factors such as the quality and quantity of labeled training data, the overall type and complexity of the model, and the application domain. A second axis is the efficiency of the implementation of a learning system on given hardware. Achieving improved efficiency—in terms of power, compute, memory usage, and quantity of data required for training—is broadly essential for scientific applications. Included in efficient implementations is the use and impact of reduced-precision hardware offered in current hardware, and novel computing hardware (quantum, neuromorphic) and associated programming paradigms in future platforms (see Chapter 13, Hardware Architectures).

While there have been significant improvements in and variants of training algorithms, the grand challenge of an efficient, general-purpose algorithm for learning remains unsolved. Further, nested nonlinear ensembles of linear models are undoubtedly not the last great learning architecture that will emerge;

novel model forms may exhibit profound advantages in terms of data efficiency.

In addition to the general learning problem, significant challenges remain for specific classes of AI models. For example, AI-based control systems rely on semi-supervised and reinforcement learning, which are inefficient, produce "brittle" systems, and are non-transferable. Efficient continuous learning systems that handle data streams at the edge and remain validated must be developed. And it will be necessary to rethink the learning process—and artificial reasoning in general—for systems, including approximate, neuromorphic, and probabilistic computing, to make them computationally tractable for many real-world problems. Further study of human neural systems and the learning process may yield significant insights, new abstractions, and complexity classes beyond those conventionally in use at present.

3. Advances in the Next Decade

Many opportunities exist to advance the foundational building blocks of AI over the next decade. We highlight a few of the areas where mathematical, statistical, and information-theoretic advances are required to address the above grand challenges. These advances entail the development of new algorithms, theory, and modeling paradigms.

Exploiting scientific knowledge. Approaches to leveraging domain knowledge include using custom loss functions; selecting decision or latent variables; applying physical constraints (e.g., conservation laws); leveraging Bayesian or probabilistic graphical models; using simulations to augment or generate training data; and exploiting known smoothness, sparsity, or other low-dimensional structures. Many of these approaches have been tested in particular areas of ML, but mathematical advances are required to establish principled ways for the incorporation of domain knowledge throughout AI and to understand the induced tradeoffs. Each of these approaches has limitations and requires significant foundational research.

Creation of surrogates. AI presents a unique opportunity for creating data-driven surrogate models that are potentially orders of magnitude faster to run than first-principles simulation codes and can be particularly effective in the ability to simulate physical processes that span many spatial and temporal scales. Some of the unique challenges for AI systems revolve around a careful characterization of the generalization limits, proofs of interpolation/extrapolation, robustness, assessment of confidence associated with predictions, and effects of the input data. Rigorously understanding these tradeoffs will impact not only model selection in AI systems, but also the creation and investigation of new classes and types of models.

Numerical optimization. Optimization algorithms, differentiation techniques, and models form the foundation of training in AI. Both the loss landscape of these models and the traversal of this landscape by algorithms are poorly understood. There is a significant opportunity to improve understanding about the effect of incorporating domain knowledge in the form of constraints or regularization terms. How do these approaches affect the solution manifold and the ability of fast algorithms to consistently find this manifold? What principles about network and model selection does this inform? What accuracy is needed in derivative and loss evaluations? What guarantees of optimality can be established? Opportunities exist for fundamental advances in this area to impact AI for science from the HPC facility to the edge.

Uncertainty quantification (UQ). An important aspect in the development and application of AI is the quantification of uncertainties. Where AI and ML are used in physics-based applications, established approaches to UQ are applicable. In other cases, particularly in classification problems, ML models tend to be highly nonlinear systems that are extremely

sensitive to input data, and small (e.g., undetectable to the human eye) changes can lead to misclassification. Several approaches to dealing with uncertainty (e.g., Bayesian neural networks) are computationally intractable for many AI problems; significant expansion of these approaches or new, more efficient alternatives are needed. Known and emerging UQ techniques can also be used to detect overfitting and select the simplest possible model.

Graph-based ML and AI. Graphs arise naturally in many scientific domains (e.g., molecules, protein interaction networks, community networks). Structuring data and knowledge representations in terms of graphs and exploiting the topology information available from a graph representation can be critical to realizing tractable algorithms and obtaining better outcomes in tasks such as classification, clustering, and prediction of missing data. Important questions need to be addressed, including, what is the most relevant graph representation obtainable from noisy data for a problem and how can it be computed efficiently? How can the topological information available in graphs be best exploited within an AI model? How can the time complexity of AI computations involving massive graphs be tamed? How can algorithms be adapted to work with dynamic graphs, and how can streaming algorithms be designed when the graph cannot be stored?

Data/model fusion and representation. Current AI and ML systems tend to analyze one type (mode) of data. However, most physical systems include data of different types or modes. For example, environmental sensor input must be combined with video streams for effective control of manufacturing processes; audio and text input must be combined in sentiment analysis; and multispectral sensors must be incorporated into environmental monitoring systems. The different modes often have fundamentally different characteristics and represent different types of information. This leads to challenges in fusing data and models across the different modes, such as the encoding and representation of knowledge or events in ways that allow an AI model to establish correlations across the different modes, and the transferability of knowledge from one mode to enable more efficient learning in other modes. Furthermore, DOE is unique in the breadth of diverse datasets and representations produced by various simulations, experimental and observational devices, and computing and networking facilities. Developing and applying AI methods successfully will require that abstractions and algorithms are aware of and target the intrinsic properties of datasets and representations (e.g., big vs. small, structured vs. semi-structured vs. unstructured, sparse vs. dense, space vs. time vs. space-time, graphs, noisy/missing/mislabeled data, multi-variate/-physics/-scale/-modal) to achieve optimal results.

Interpretable and explainable AI. While the ultimate goal of AI research may be fully autonomous systems and artificial general intelligence, the larger potential for the near future is augmenting human intelligence—including, for example, accelerated scientific discovery and engineering design, engineered safety systems, and improved medical diagnoses. In the context of accelerating science and engineering, as AI methods make inroads and produce state-of-the-art results for data analytics, surrogate modeling, inverse design, and control applications, advanced capabilities are needed to explain the behavior of AI models and to map the internal representation of AI models to domain-specific concepts.

Hypothesis generation, design of experiment, and causal analysis. Validation is the process of determining whether an AI model is appropriate for the application or decision for which it is being used. One of the fundamental questions during validation is whether the AI model is making the right decision for the right reason. For example, has the AI model learned spurious correlations, or can the model

determine the control variables? In short, can AI be used to identify causal variables or distinguish between cause and effect? Typically this cannot be done with a single training dataset. Instead, the AI model needs to be trained to construct a hypothesis, typically a counterfactual one, and to design an experiment—including the collection of data (and the suitability of that data)—to test that hypothesis.

Robustness/stability. Robustness generally refers to an algorithm's ability to deal with errors in the input data or errors during execution of a program. This also includes the ability of AI to withstand an adversarial attack, as well as the ability to deal with corner cases and rare events that may not appear in the training data. Similarly, stability refers to the ability to deal with rounding and other errors that are an intrinsic part of any numerical algorithm. In many cases, classical numerical analysis approaches can quantify and control these errors; however, foundational research adapting these results to ML algorithms and developing AI-specific approaches to improving the robustness and stability is required. Identifying the limits of AI methods and models—for example, in terms of input or training data ranges beyond which errors can grow undesirably—would advance understanding.

Reinforcement learning (RL) and beyond. RL forms the foundation of most AI-based control and policy systems. RL is the process of teaching an AI model to take actions based on a current state, an environment, and a reward function; it has been studied historically in the context of dynamic programming, Markov decision processes, and control theory. Within the context of AI, RL has been used successfully in many applications, most visibly by DeepMind for player policies for increasingly complex games. Despite RL's recent success, many challenges must be addressed for scientific and engineering control applications. For example, action/reward shaping for control decisions may lead to computationally inefficient and costly training; there is a tradeoff between exploitation and exploration, which is similar to the tradeoff between depth-first and breadth-first search; and there is often a narrow applicability regime and a lack of robustness in training control systems. The human-computer interface and explainability must also be considered in the context of RL-based control systems.

Real-time learning and control. AI impacts are typically attributed to the availability of both data and computing. However, in some science applications one can see the dual problems of too much data and too little computing and storage. In such cases, one will not be able to store even a small fraction of the generated data, nor will one have the ability to (re)train models from scratch. Furthermore, the data may have low information content and may end up corrupting models if used incorrectly. Even today, the cost of training a single, albeit large, deep neural network has been estimated in the tens or hundreds of millions of dollars of power and computing capacity. Improving the ability to train an AI model continuously (e.g., with streaming data that is discarded immediately after use) and to deploy the model in real-time requires advances in areas such as adaptive models, event and anomaly detection, transfer learning, plasticity, and validation.

Unsupervised learning and dimension reduction. Much of the data used in scientific and engineering ML is unlabeled. For example, the scientific objective may be to identify patterns in datasets, find clusters, estimate distributions, compress data, identify latent variables, or reduce the dimension of a large dataset. First-principles simulations can be used to offset partially the lack of labeled data (e.g., through the use of simulated data for training, generative adversarial networks [GANs], or direct incorporation of physical laws). Advances will depend on continuing research in areas such as matrix factorizations, kernel methods, GANs, and autoencoders, with a particular focus on incorporating physical knowledge and explainability (e.g., in the

determination of latent variables and other lower-dimensional representations).

4. Accelerating Development

Data and models are growing at an unprecedented scale. A business-as-usual approach for funding research on the foundations of AI for science is insufficient for staying ahead of this deluge, let alone to transform such data and models for scientific understanding. The use of AI in scientific and engineering applications is often constrained by the lack of good and labeled data; the inefficient, brittle, and unpredictable training of AI models; and the lack of assurance, including UQ, validation, and interpretability. Key investments to accelerate development along the above advances include the following:

The use of scientific principles, modeling and simulation, and domain-specific knowledge to inform and advance AI. Focusing here would spur the ability to learn effectively with orders of magnitude less data and/or to use the same data for otherwise unthinkable predictive power and generalizability.

Addressing robustness, uncertainty quantification, and interpretability of AI systems. Increased understanding of the sensitivities and limitations of AI models and improving scientists' ability to interpret AI outcomes would significantly accelerate the adoption of AI as a scientific capability.

Learning for inverse problems and design of experiments. Inverting traditional cause-to-effect models to learn what causes could have produced an effect, and then to efficiently generate experimental campaigns to test these hypotheses, would broaden the scientific method.

Reinforcement and active learning to develop AI for control and data acquisition systems. Advances to directly address dynamic operations and real-time feedback scenarios would narrow the distance from the AI to the instrument, detector, lab, and facility.

5. Expected Outcomes

A research agenda supporting algorithmic and theoretical advances in AL and ML will have a profound impact on science, society, and industry. Successfully addressing the challenges identified above will reap huge rewards and enable rapid progress in areas such as advanced manufacturing, energy distribution and generation, mobility and transportation infrastructure, bioenergy, health science, and advanced materials design and synthesis.

Primary outcomes of advancing the foundations of AI will be to maximize the understanding realized from science-informed AI, to increase trust in ML and AI as scientific techniques, and to provide efficient computational algorithms—for implementations in diverse and heterogeneous computing and instrument hardware—for generating these models.

With these advances, we expect that AI and ML will become accepted and well-characterized tools in the modern scientific computing toolbox, and the abstract models generated are understood for use in a variety of tasks. Minimizing the risks associated with AI uses is especially important in high-consequence applications. Increased trust will also further the adoption of AI and embedded intelligence in everything from edge devices to networks to HPC facilities. Significant improvement in the efficiency of ML will enable more accurate surrogate models of complex physical systems (e.g., reacting flows or failure mechanisms in materials), optimization algorithms for inverse problems in materials characterization and design, and more accurate computation uncertainties necessary in all science and engineering disciplines.

6. References

1. Thomas, N., et al. Tensor Field Networks: Rotation- and Translation-Equivariant Neural Networks for 3D Point Clouds. Arxiv Preprint, arXiv:1802.08219, 2018.
2. Jordan, M. I., Artificial Intelligence: The Revolution Hasn't Happened Yet. *Harvard Data Science Review* (2019). doi:10.1162/99608f92.f06c6e61.
3. Baker, N., et al. *Workshop Report on Basic Research Needs for Scientific Machine Learning: Core Technologies for Artificial Intelligence*, 2019. doi:10.2172/1478744
4. Arridge, S., Maass, P., Öktem, O., and Schönlieb, C. Solving Inverse Problems Using Data-Driven Models. *Acta Numerica,* **28**, 1-174. doi:10.1017/S0962492919000059.
5. Kondor, R., Trivedi, S. *On the Generalization of Equivariance and Convolution in Neural Networks to the Action of Compact Groups*. Proceedings of the 35th International Conference on Machine Learning, PMLR 80:2747–2755, 2018.
6. Goodfellow, I., et al. *Generative Adversarial Nets. Advances in Neural Information Processing Systems*, 2014.
7. Chen, X., et al. Infogan: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. *Advances in Neural Information Processing Systems,* 2172–2180 (2016).
8. Arjovsky, M., Chintala, S. and Bottou, L. *Wasserstein Generative Adversarial Networks.* International Conference on Machine Learning (pp. 214–223, 2017).
9. Paganini, M., de Oliveira, L. and Nachman, B. CaloGAN: Simulating 3D High Energy Particle Showers in Multilayer Electromagnetic Calorimeters with Generative Adversarial Networks. *Physical Review D* **97**: 014021 (2018).
10. Zhang, K., Zuo, W., Chen, Y., Meng, D. and Zhang, L. Beyond a Gaussian denoiser: Residual Learning of Deep CNN for Image Denoising. *IEEE Transactions on Image Processing,* **26:** 3142–3155 (2017).
11. He, K., Zhang, X., Ren, S. and Sun, J. Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 770–778, 2016.
12. Bottou, L., Curtis, F.E. and Nocedal, J. Optimization Methods for Large-Scale Machine Learning. *Siam Review,* **60:** 223–311 (2018).
13. LeCun, Y.A., Bottou, L., Orr, G.B. and Müller, K.R. *Efficient Backprop. Neural Networks: Tricks of the Trade*, Springer, Berlin, Heidelberg, 2012.
14. Sutskever, I., Martens, J., Dahl, G. and Hinton, G. On the Importance of Initialization and Momentum in Deep Learning. *International Conference on Machine Learning,* 1139–1147 (2013).
15. Duchi, J., Hazan, E., and Singer, Y. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 2121–2159 (2011).
16. National Research Council.– *Frontiers in Massive Data Analysis* Washington, DC: The National Academies Press. 2013. https://doi.org/10.17226/18374.
17. Mustafa, M., et al. CosmoGAN: Creating High-Fidelity Weak Lensing Convergence Maps Using Generative Adversarial Networks, *Computational Astrophysics and Cosmology* **6**, (2019).
18. Wu, J. L., et al. *Enforcing Statistical Constraints Generative Adversarial Networks for Modeling Chaotic Dynamical Systems*, Cornell University, 2019. https://arxiv.org/abs/1905.06841.

19. Raissi, M., Perdikaris, P., Karniadakis, *G. E. Physics Informed Deep Learning (Part I): Data-Driven Solutions of Nonlinear Partial Differential Equations*. https://arxiv.org/abs/1711.10561.

20. Yang, L., et al. Highly Scalable, Physics-Informed GANs for Learning Solutions of Stochastic PDEs (SC'19 Deep Learning on Supercomputers Workshop).

21. Weiler, M., Geiger, M., Welling, M., Boomsma, W. and Cohen, T. *3D Steerable CNNS: Learning Rotationally Equivariant Features in Volumetric Data.* Advances in Neural Information Processing Systems, 10381–10392, 2018.

22. T. D. Bui, S. Ravi and V. Ramavijjala, *Neural Graph Learning: Training Neural Networks Using Graphs*, Proceedings of 11th ACM International Conference on Web Search and Data Mining, 2018.

23. R. L. Murphy, B. Srinivasan, V. Rao and B. Ribeiro, Relational Pooling for Graph Representations, Arxiv:1903.02541, 2019.

24. K. Xu, W. Hu, J. Leskovec and S. Jegelka, How Powerful are Graph Neural Networks? ArXiv:1810.00826v3, 2019.

25. Tschannen, M., Bachem, O. and Lucic, M., 2018. Recent Advances in Autoencoder-Based Representation Learning. arXiv preprint arXiv:1812.05069.

26. Ben-David, S., Hrubeš, P., Moran, S. et al. Learnability Can be Undecidable*. Nat Mach Intell* **1:** 44–48 (2019). doi:10.1038/s42256-018-0002-3

27. Tshitoyan, V., et al*.* Unsupervised Word Embeddings Capture Latent Knowledge from Materials Science Literature. *Nature* **571**, 95–98 (2019). doi:10.1038/s41586-019-1335-8

28. Swain, M. C. and Cole, J. M., 2016. ChemDataExtractor: a Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *Journal of Chemical Information and Modeling*, **56**: 1894–1904 (2016).

29. Clark, P., et al., 2019. From 'F' to 'A' on the NY Regents Science Exams: An Overview of the Aristo Project*.* arXiv preprint arXiv:1909.
01958.

This page intentionally blank.

11. Software Environments and Software Research

The DOE Office of Science has an opportunity and need to research and develop software to address the office's research mission. Such an effort would complement large investments by industry to develop AI software environments. The DOE has deep expertise in simulation, modeling, and large-scale data analysis, and it also operates the largest and broadest set of user facilities for experimental and observational science, including light sources, telescopes, and genomics facilities that have growing computing and data-analysis requirements (see also Chapter 16, Facilities Integration and AI Ecosystem). There is an urgent need to develop software and computing environments that enable AI capabilities to be seamlessly integrated with large-scale HPC models and the growing data-analysis requirements of experimental facilities.

1. State of the Art

There is currently a proliferation of software and frameworks for data analysis and machine learning. Top deep learning and ML frameworks today include scikit-learn, TensorFlow, PyTorch, and Keras, but new software and frameworks are being released regularly. These new frameworks are primarily developed and led by industry, with some notable contributions from academia for software such as Spark and Jupyter. The software is open source, though not open governance, and is often controlled and sponsored by industry leaders, such as Google and Facebook.

There are a few notable gaps between state-of-the-art and DOE scientific requirements when it comes to software for AI. First, DOE researchers produce massive amounts of data from simulations and models that can benefit from the integration of AI capabilities. These are often challenging datasets with multidimensional data and can also include nonimage-based data. Second, DOE runs unique user facilities that produce petabytes of data, have no counterpart in industry, and require new AI software and capabilities. Finally, many of the DOE scientific datasets need the scale of HPC systems for analysis, and those systems can have unique architectural features that require software attention and investment, such as large-scale I/O subsystems and heterogeneous compute elements. With DOE's challenging datasets and deep expertise in data analytics, simulation, and modeling, DOE researchers are well positioned to contribute unique enhancements to the AI software stack.

2. Major (Grand) Challenges

When considering the impact of AI on software environments and software research, three significant opportunities are apparent. First, the integration of AI into the "inner loop" can lead to more effective simulations (see also Chapter 10, AI Foundations and Open Problems). For example, leveraging AI within a simulation could lead to more efficient modeling by virtue of the development of digital twins during runtime. Second, integration of AI into the analysis approach could lead to faster generation of analytical results, automate the identification of anomalous behavior, and ultimately lead to automatic hypothesis generation. Finally, the integration of AI into the management and control of research labs, facilities, experiments, and workflows (i.e., the "outer loop") can help achieve a variety of goals. Examples include adapting workflows in response to new hypotheses generated during the workflow, scheduling resources for more efficient use of facility hardware, and dramatically reducing the total cost of operating facilities. These three grand challenges are not orthogonal and would provide the greatest impact when examined together (see also Chapter 16, Facilities Integration and AI Ecosystem).

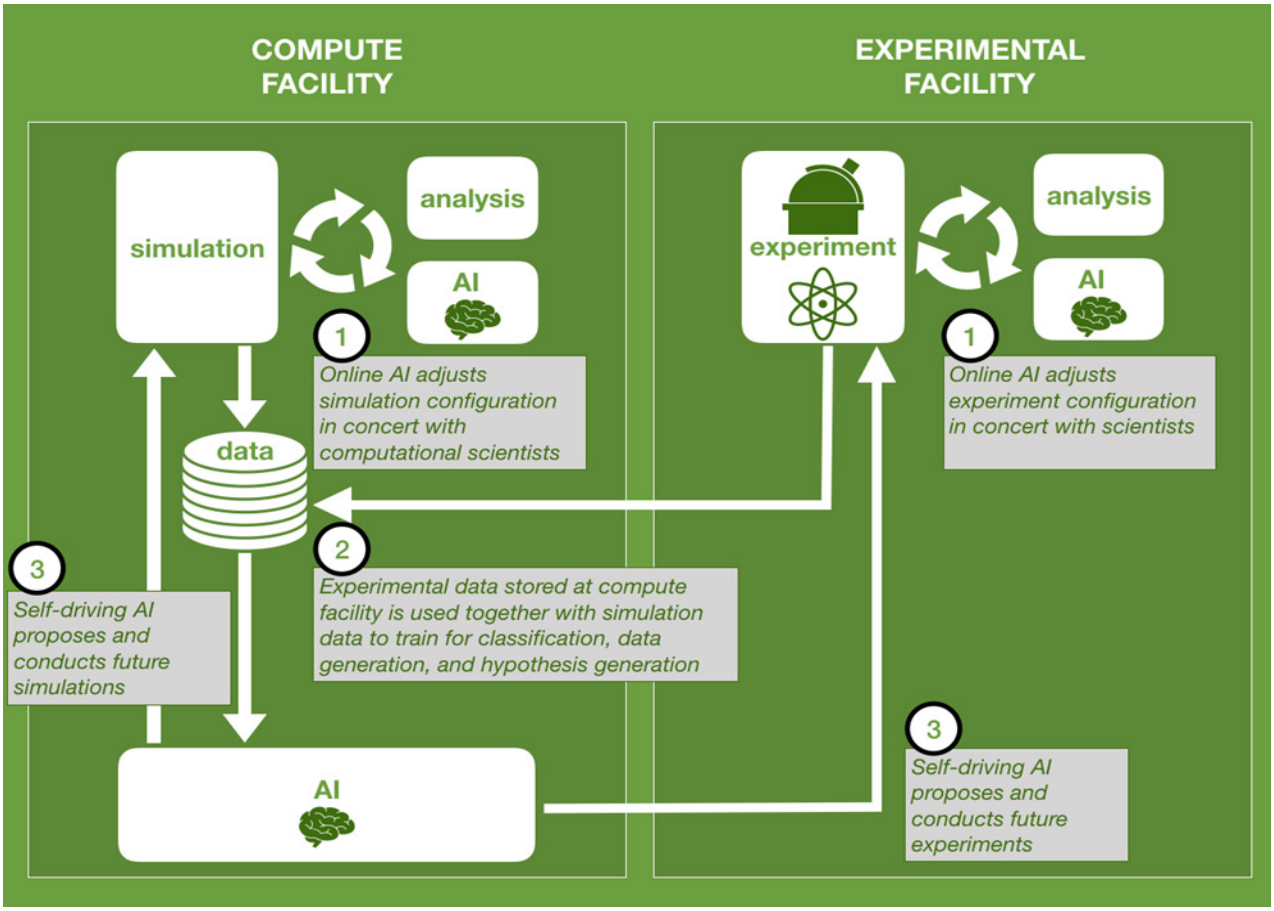


Figure 11.1 *Three opportunities for the integration of AI into software environments have the potential for dramatic impact on DOE science: (1) Within the "inner loop" of simulations and experiments, (2) to accelerate and enhance traditional analysis approaches, and (3) in the "outer loop" to assist in the management and control of workflows, laboratories, and facilities.*

Develop software for seamless integration of simulations and AI. DOE is the premier agency for large-scale simulation and modeling of physical phenomena because it has deep institutional knowledge and expertise in numerical methods, solvers, and parallel implementations. There is an opportunity to improve the performance, efficiency, and fidelity of traditional simulations by integrating AI capabilities. Such a system would allow the integration of data from different sources, in different formats, and over different time domains into existing mathematical models and adapt in real time to changing model conditions. In addition, AI model-generated data can be validated against in-memory simulation data; by comparing results from *in situ* analyses on simulation-generated and model-generated data, one can also determine thresholds at which the model-generated data are sufficiently accurate and, therefore, determine when the trained model can replace the simulation kernel. Similarly, AI approaches could be employed to aid in mapping simulation workflows onto upcoming complex and heterogeneous platforms, revising the use of resources over the course of workflow execution through increasingly refined and accurate performance models. These approaches have the potential to significantly impact traditional simulation and modeling by improving the performance of simulations [1].

This would lead to a new hybrid computation model, combining traditional simulation with AI results in a model that runs more efficiently or produces higher fidelity results. For example, a traditional mathematics-based climate model (i.e., a multiscale, multiphysics simulation) could be enhanced by replacing a

computationally intensive kernel with stochastic properties with a physics-informed ML approach. Alternatively, an AI system could learn the representations produced by a simulation kernel, and then the kernel could be replaced with a better-performing, lower-complexity generative model. Furthermore, the output of this model could be combined with ML hydrology models, flooding maps, and evacuation routes, allowing new and more accurate predictions (see also Chapter 2, Earth and Environmental Sciences) [2].

Significant investments need to be made in software and programming environments to realize this vision. Today, the coupling of traditional modeling and simulation codes with AI capabilities is largely a one-off capability, replicated with each experiment. The frameworks, software, and data structures are distinct, and APIs do not exist that would enable even simple coupling of simulation and modeling codes with AI libraries and frameworks. *In situ* data analysis requiring ML capabilities suffers from the same limitations. Significant software engineering investments are needed to enable reusability and composability that would reduce the integration overhead between simulations, data analysis, and AI, along with the integration of new foundational research advances into AI software. This includes addressing the need for composable data structures and modular elements that enable seamless movement between simulation, data analysis, and AI algorithms, as well as improvements in performance modeling and programmatic control of task placement in workflow systems to enable autonomous mapping of tasks to heterogeneous resources at runtime.

In addition, at present, the parameters of models and the choice of solvers is largely determined by human expertise and is fixed at compile or runtime [1]. An integrated AI and simulation software environment would enable a model to use one method in a given time step and a different one in the next for a more optimal or faster converging system. To transition to a mode where simulations can adapt in real time, however, investments are needed in areas such as enabling real-time annotations and descriptions of schemas to allow the real-time adaptation of models and analysis. Computer scientists and software developers cannot do this on their own; it is essential that software capabilities be co-designed in concert with algorithm developers and domain science experts. One sometimes-overlooked facet of this challenge is the corresponding need for enhancements in data storage, access, and management that would facilitate rapid identification of relevant data, transformations between different data representations, and capture of relevant provenance to assist in reproducibility of results (see also Chapter 12, Data Life Cycle and Infrastructure).

Develop software for knowledge extraction and hypothesis generation. The volume of data, and the knowledge that can be derived from data, is expanding exponentially in nearly every area of science. Creating next-generation AI software that identifies gaps in existing knowledge and relevant data can enable the generation of new scientific hypotheses relevant to each scientific question and provide recommendations for knowledge discovery. Such intelligent software and frameworks will be able to investigate various possibilities, parameters, and models in a scalable manner to gain fundamentally new insights in specific science domains. In addition, as interdisciplinary research is gaining momentum, DOE is well-positioned with its breadth of both scientific and foundational interests to serve as the nexus for this work. AI-enabled software can use meta-learning techniques to identify potential overlaps across the different science domains and generate hypotheses that can lead to new discoveries. Such AI software can keep track of many disparate but relevant data points within and between different sciences as well as suggest next steps.

For example, experiments have traditionally been designed and conducted by humans, with the help of computational simulations and analyses to identify and constrain the design. This cycle of experimentation benefits from the body of data amassed in previous experiments. AI can provide further insights when trained on experimental data, combined with simulation data and analysis results, culminating in a more precise representation of the phenomena being studied that incorporates physical constraints, domain knowledge, and human expertise. On the basis of a continuously growing collection of validation data from experiment and simulation, the predictive power of AI models will improve over time; by identifying areas where the AI models fall short, one could call for more data and experiments to improve performance in the low-performing context. As these experiments are conducted and the model learns from the resulting data, its predictive power will improve in the poor areas, and it will become better able to generate hypotheses for subsequent experiments (see Chapter 4, High Energy Physics and Chapter 14, AI for Imaging).

One of the major challenges in enabling knowledge discovery and hypothesis generation is the reuse of existing and future data. Both the scale and potential disparate modality of scientific data, be it simulation output or experimental observations, are unique when compared to other traditional, nonscientific AI training datasets. The amount of data, and the existence of different data types and models, requires AI software that can enable interoperability and knowledge extraction by reusing the data from different domains. The use of natural language processing in AI training will be of growing importance to integrate across the disparate data modalities, which may include scientific literature in addition to experimental and simulation data. Another critical challenge is to generate the right hypothesis, as research across all science domains has become incredibly complex and it is extremely hard to connect the relevant data points. Existing ML techniques do not provide substantial understandability of the models and the outcomes. Finally, scalability and performance will be a major challenge in knowledge discovery. The sheer amount of data and associated variables across different science domains need a scalable framework that can run HPC systems.

As researchers identify gaps in data and generate new hypotheses, significant investments are needed for developing intelligent, scalable AI software frameworks that can leverage existing data, models, and the associated provenance about the training and analysis methods. Such a framework would provide real-time recommendations for understanding the data gaps and exploring the hypotheses. Hence, next-generation AI software needs a self-improving metadata layer that can continuously learn from the data and models to enable algorithmic discovery from data. Such AI software would use the provenance and metadata to describe the model architecture, parameters, and data. Investments are also needed to identify the unique architectures that cannot only help determine the right network architecture for a particular science domain, but also cross multiple domains. This will require additional investments in infrastructure for sharing the knowledge and associated data.

Enable self-driving experiments with AI integration and controls. Computing has become increasingly pervasive as a tool at experimental science facilities, from simulating phenomena to controlling systems, analyzing experiment data, and driving hypotheses to explore in subsequent experiments (see Chapter 4, High Energy Physics and Chapter 14, AI for Imaging). AI can be regarded as an embodiment of this process, touching all of these aspects and driving the hypothesis-simulation-analysis cycle as a whole.

There is a wide range of experimental science projects, and their integration with compute and data capabilities is varied; however, the

direction a number of experiments are moving toward is more frequent online analysis and adaptation of experiments. For example, scientists at certain light sources use analysis of imaging for decision-making in near real time. These analyses are typically run at the experimental facilities within the constraints of time-to-solution and compute availability. In addition to employing AI in analysis and hypothesis generation as described above, AI could be used to act on these results, adapting to data as they emerge by adjusting the parameters of the experiment toward specific goals, such as protecting resources, maximizing the data gathered related to a specific phenomenon, or following up on surprising or anomalous results. By automating high-level decision-making, experiments could proceed without scientists onsite, and scientists would be better able to focus on high-level goals of the discovery process rather than directly monitoring individual experiments. Ultimately, this AI capability could also identify experiments that cannot be executed with current devices but are likely to uncover promising results, pointing toward promising new experimental capabilities.

More generally, complex workflows are an integral part of scientific discovery, and increasingly these workflows are defined programmatically so that they may be executed as an integrated system. Just as in traditional programming, handling the wide variety of possible outcomes from specific tasks is tedious and error-prone, leading to workflows that often terminate in the face of unexpected results. By allowing scientists to describe workflows in terms of high-level goals, building-block tasks (i.e., experiments, simulations), and rough models of the costs of those tasks, an AI system could instead generate a specific workflow, incorporating expert knowledge, to accomplish those tasks, adapting as results are uncovered or new data become available and refining the models of costs (e.g., in time). Similarly, AI can be integrated into the management of experimental controls and computational resources. An AI system could be allowed to observe the incoming stream of application workflows, the state of the experiment, computational and storage resources, and the behavior of applications that have been executed, adjusting the allocation of resources toward goals such as maximizing utilization of the platform; enabling rapid execution of priority jobs (e.g., in response to a national or regional emergency); or throttling power utilization in times of high demand (e.g., in response to a regional heat wave). By integrating AI into the management of these systems, the need for immediate responses from facility staff would be reduced, freeing resources to better assist application teams in making the best use of these resources. (See Chapter 15, AI at the Edge for further descriptions of edge computing use cases.)

Numerous software advancements are needed to achieve this vision of AI-driven science. Workflow description capabilities need to be enhanced to allow for expert knowledge, goals, building-block tasks, and constraints (e.g., safeguards) to be described in a manner that can be used to automatically construct workflows. Telescoping language approaches, for example, might be beneficial in this context. Additionally, new tools for programmatically describing relationships are needed to enable AI systems to reason about cause and effect in these systems, or at least to provide a starting point on which improved models can be built. Finally, significant investments are needed in systems that enable programmatic control of instruments.

3. Advances in the Next Decade

The current rapid pace of development in ML methods is a direct consequence of the availability of open-source software frameworks, such as PyTorch and Tensorflow, that tightly integrate algorithmic and programming techniques (e.g., optimization and automatic differentiation) with modern hardware. They lower the barrier for entry and enable rapid iteration on the new domain-specific ML architectures.

Extrapolating this development trend, we can expect a software stack that facilitates efficient use of a broader range of algorithmic and mathematical techniques, making it as easy to use methods from geometry, topology, functional analysis, optimal transport, and constraint satisfaction as it is today to use differentiable programming. Simultaneously, given the current trends, data sizes, and the role of hardware in ML, it is reasonable to expect the evolved software stack to take even greater advantage of the new hardware accelerators, as well as to target distributed computing architectures (see also Chapter 13, Hardware Architectures and Chapter 16, Facilities Integration and AI Ecosystem). This presents a potential danger of more fragmentation into proprietary silos, as well as targeting architectures like the industrial "cloud," rather than DOE supercomputers.

There is also a clear recognition in the industry of the challenges of data and workflow management associated with the vast volumes of training data and complicated processing pipelines. There are numerous efforts to automate and standardize approaches to these challenges, and they are likely to bear fruit in the coming decade. It is important to note that these industry efforts are driven by data types that are often very different in terms of modality, dimensionality, resolution, and scale from those produced by DOE experimental, observational, and computational facilities.

4. Accelerating Development

Three early activities would help put efforts in software environments on track for success. First, a strong gap-analysis effort that identifies internal requirements and assesses existing tool capabilities is critical to understanding where investments are most needed. Initial work in this direction has been performed in specific science domains already, and reports are being written.

Second, it is important for computer scientists, applied mathematicians, and domain scientists to work together to co-design solutions that integrate AI into these complex scientific endeavors. These types of partnerships have been successfully established toward other applications of computer science and applied mathematical techniques, and in many cases these partnerships can be launchpads for new AI-focused efforts. Early experiences in these partnerships will further inform research investments as well.

Finally, looking outward, it is critical to understand how AI and associated technologies being developed outside the DOE are managed and governed to assess whether these tools will be viable for DOE use over the long term. If our strategy involves heavily leveraging AI software technologies from other sectors, then DOE must engage with these communities and establish itself as a contributor to help ensure the relevance and effectiveness of these packages on HPC systems.

5. Expected Outcomes

Advanced and capable software is indispensable for scientific discovery through simulation, modeling, and data analysis. With AI radically transforming numerous fields, it is crucial that investments are made in software to support new AI capabilities in support of the DOE mission. The integration of AI into traditional simulation and modeling will improve performance, efficiency, and fidelity of models of complex phenomena, and the ability to integrate models with historical and real-time data will improve the predictive accuracy of systems. Next-generation AI software frameworks will automatically identify gaps in existing knowledge and relevant data and explore new and unexpected scientific hypotheses, leading to potentially ground-breaking discoveries already within reach of

DOE experiments. Experiments and workflows with AI augmentation will optimize use of premiere DOE computing and experimental facilities and identify key features of these facilities that lead to even more effective future platforms. Investments in software for AI will result in software artifacts, new or enhanced frameworks, models, and libraries that will broadly benefit the DOE user community.

6. References

1. Baker, N., et al. Workshop Report on Basic Research Needs for Scientific Machine Learning: Core Technologies for Artificial Intelligence, DOE Office of Science Technical Report, 2019.
2. Gil, Y. & Selman, B., A 20-Year Community Roadmap for Artificial Intelligence Research in the US, 2019.

This page intentionally blank.

12. Data Life Cycle and Infrastructure

Much recent progress in AI has been fueled by the availability of massive data. For example, dramatic progress in deep neural networks for image understanding owes much to the ImageNet database of more than 14 million annotated and labeled images. Science, too, is about data, and the AI-driven transformation of science will require major changes in data generation, organization, processing, and sharing. This section reviews these changes and the research and development necessary to support this vision.

Consider the following scenario: *It is 2030. DOE scientists are working to develop a low-cost, high-performance solid-state battery for use in vehicles. Intuiting that disordered materials holds promise, they task an AI system with identifying candidate formulations. Informed by 400 years of physics knowledge, 100 years of scientific literature, and 40 years of experimental data from DOE labs, universities, and industrial collaborators, the AI system is able to evaluate options faster than any human expert. It suggests new families of disordered materials that may have acceptable stabilities, power densities, and manufacturing costs. However, it also shows high uncertainties in its predictions.*

To collect more data, the scientists task the AI system with defining and running a series of experimental and simulation studies in new autonomous laboratories and on postexascale conventional and quantum computing systems. New data integrated into the AI model motivate further experiments. Within weeks, the human expert/AI team has refined understanding to the point where large-scale manufacturing can be considered. Provenance information collected throughout allow for reuse and meta-analysis of discovery processes.

Central to this scenario is the existence of a large, well-curated, and integrated collection of data of many types—from point measurements to massive video—and from many sources, including the scientific literature (e.g., Chapter 1, Chemistry, Materials, and Nanoscience), experiments, simulations, and vehicle fleets, and encompassing both public and proprietary elements. Each item within this data collection is documented with details as to where, when, and how it was generated. Furthermore, the data collection accommodates dynamic additions as new knowledge is created.

Such data collections do not exist today, outside of a few narrow domains. Laboratories are not, in general, set up to preserve data. Many data are recorded in archaic formats and media without annotation. Descriptive metadata are inadequate and inconsistent. Data are rarely findable, accessible, interoperable, or reusable (FAIR) [1], whether by scientists or by AI systems. Data collections are often biased by a tendency not to publish negative results (i.e., the “file drawer problem”). Autonomous laboratories that can generate data at scale and under AI direction exist only in prototype forms.

AI-driven discovery across the broad range of domains important to DOE science will require transformations in both the methods and infrastructure used to acquire, organize, and process data and the policies that govern data access. These advancements must proceed via a process of co-design, with progress in methods informing infrastructure and policy changes and vice versa. The ultimate goal is a system of methods and infrastructure that enables the coordinated creation, application, and update of large quantities of data and knowledge as well as associated models, workflows, computations, and experiments (Figure 12.1).

This chapter makes the case for three priority research directions, or grand challenges, to produce the methodological advances required to create AI-ready data infrastructure.

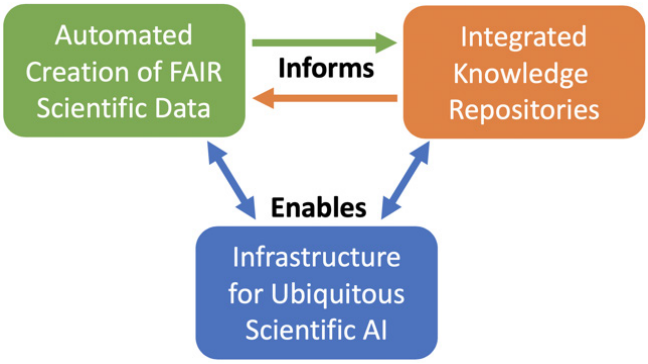


Figure 12.1 *AI-driven science requires simultaneous advancements in the methods, infrastructure, and policies used to acquire large-scale scientific data, integrate data and symbolic knowledge, and structure data infrastructure.*

Automate the large-scale creation of FAIR scientific data. Given data's central role in AI-driven science, new technologies, methods, and best practices are needed to scale the generation, capture, annotation, and organization of data from experiments, observations, and simulations to produce large collections of FAIR data for AI-enabled discovery.

Integrate data and theory to create converged knowledge repositories. Realizing the full potential of scientific AI requires a convergence of data and symbolic representations. To this end, new methods are required to synthesize AI models from data and to integrate symbolic representations of scientific knowledge, to create knowledge collections that are similarly FAIR.

Architect new infrastructure to support ubiquitous scientific AI. As AI methods are deployed ever more widely, new infrastructural concepts and methods are required to ensure that both data and the computation required to ingest, enhance, integrate, and interpret data can be accessed efficiently and reliably—whenever, wherever, and at whatever scale required.

1. State of the Art

Despite much progress in scientific data acquisition and management, the datasets, processing methods, and infrastructure needed to engage fully in the development and application of AI methods for science are still lacking. Few communities have large collections of high-quality, curated, and labeled data suitable for use by AI systems. Even in domains where much data has been generated, silos and the lack of coordination among data collection efforts hinder broader access.

For example, in the field of materials science (Figure 12.2), data collections number in the hundreds and are distributed worldwide. The Materials Data Facility [2] indexes more than 100 data sources and operates automated data ingestion and metadata extraction pipelines to facilitate automated analyses. Nevertheless, most materials data remain unfindable, inaccessible, and noninteroperable and are rarely reused.

As a second example, the velocity at which microbiome data are generated has far outpaced current capabilities for collecting, processing, and distributing these data in an effective, uniform, and reproducible manner, even at the largest data centers. The National Microbiome Data Collaborative (NMDC) was established by the Office of Science in 2019 to build the infrastructure needed to apply consistent ontologies, annotations, and processing to create a FAIR microbiome data resource. The NMDC aims to remove roadblocks in the development of AI methods for microbiome analysis by making large quantities of labeled, curated, interoperable data available to the public. Broad success in these areas depends on overcoming challenges outlined in this chapter.

The Systems Biology Knowledge Base (kBase) [4], Earth System Grid Federation (ESGF) [5], and Atmospheric Radiation Measurement (ARM) facility [6] are further examples of DOE-supported data infrastructures that assemble large volumes of important scientific data that offer opportunities for application of AI methods.

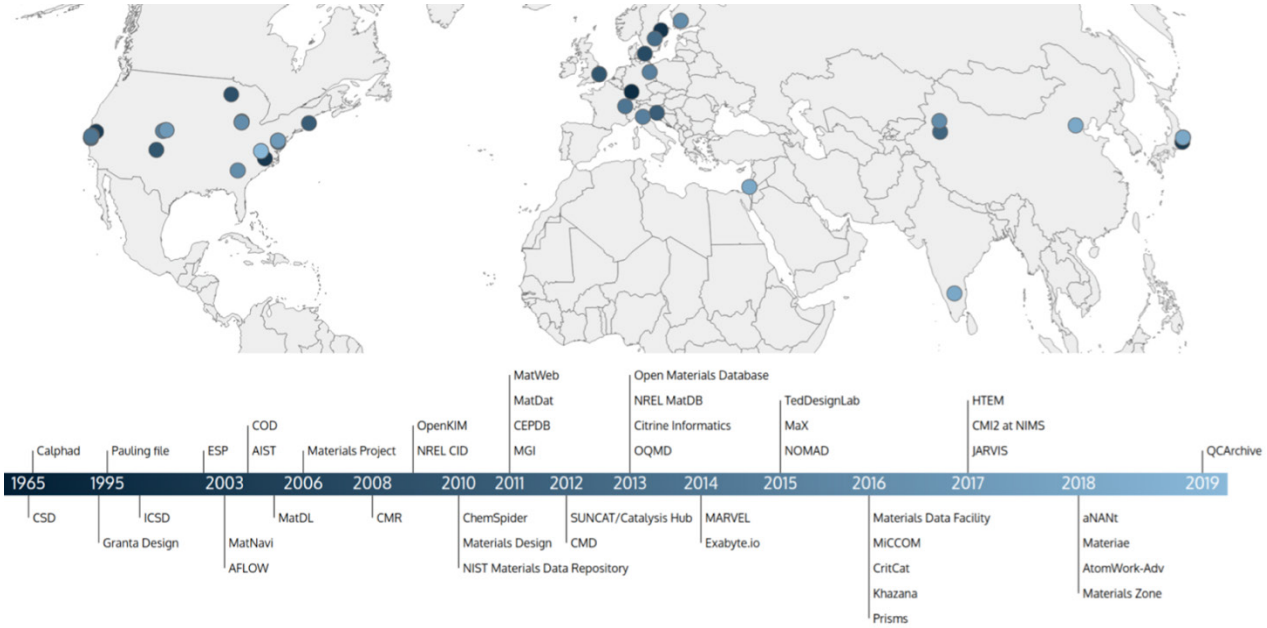


Figure 12.2 *Timeline and geographic distribution of selected materials data infrastructures and companies [3].*

Overall, the infrastructure and methods needed to enable AI methods to access, learn from, and add to a broader body of knowledge are in their infancy.

Annotation with useful metadata is an important prerequisite for widespread use of scientific data. Some communities have well-established procedures for encoding metadata in datasets, such as the climate and forecast (CF) metadata conventions used in earth and atmospheric sciences. Yet even when such conventions exist, they often fail to capture detailed annotations to support searches for specific characteristics or features within large datasets. Some recent work investigates the use of ML to generalize metadata from a subset of labeled data by classifying electron microscopy images automatically as being generated by either transmission electron or scanning transmission electron microscopy [7]. Much more work is required to streamline and simplify the process of creating metadata for scientific datasets.

2. Major (Grand) Challenges

Successful realization of AI-driven science at the scales envisioned in this report requires the creation of large collections of FAIR, AI-ready science data, and the development of methods and technologies for manipulating those data. By addressing the following grand challenges, this vision has a stronger probability of being realized.

Automate the large-scale creation of FAIR scientific data. Much scientific data today is still created laboriously through individual experiments and then organized via time-consuming and error-prone manual data acquisition, movement, and annotation steps. Many data are discarded to alleviate transfer and storage costs, and descriptive metadata are often inadequate to enable subsequent reuse. Scientists need new approaches if they are to accumulate the volume, variety, and quality of science data required for AI-driven methods. In particular, steps must be taken to automate major elements of data creation. Automation is discussed here from the perspective of data and workflows (see Chapter 11, Software Environments and Software Research). See also a recent ASCR report [8].

While harnessing existing data flows within scientific laboratories is an important first step toward creating the rich data collections needed for AI-driven science, progress will

remain limited if it is dependent on experiments defined and executed manually by human operators. High-throughput experimentation alone is not a sufficient solution, either. In many fields of science, the possible alternatives are far too numerous for exhaustive searches. Instead, autonomous laboratories are required that combine, in varying ways, high-throughput experimentation, large knowledge bases, AI methods, and human guidance to both generate data and answer questions [9,10]. The development of such autonomous laboratories will require advances in many areas, not least data management and analysis, so that AI agents can define new experiments quickly and effectively in light of extant knowledge.

Regardless of how data are generated, the automation of what are currently manual data capture and curation tasks is key to increasing the quantity of data collected and the quality and usability of those data and associated metadata, as well as supporting ontologies. Automation must support and simplify all aspects of the process, from creation to use. AI itself should be harnessed to improve this process, with autonomous data curation capabilities working to capture provenance and context information required for future reuse, and to encode associated uncertainty ranges. These new methods must be adaptable to different applications and disciplines and be able to support multimodal data collection from multiple science domains. They need to be able to organize datasets for both immediate use and subsequent reuse, without requiring scientific campaigns to plan for curation and storage independent of data gathering.

With increased automation, it can be anticipated that multimodality and diversity of data will change from being a barrier to an asset for AI purposes. To align a multitude of datasets that are collected from different sources, in different locations, at different times, and for different purposes, systems must be powerfully interoperable and able to produce data maps to guide subsequent human or AI consumers. Data will have value in unexpected ways for AI agents; when, where, and how a dataset may be used after generation are all unpredictable. Data curation decisions concerning, for example, what data to collect and what to discard, will become vital as larger amounts of science data become available. Data collections must flexibly accommodate notions of value and importance, age, and ownership to support their optimal use for AI. Furthermore, because data collection and curation decisions frequently incorporate ethics and bias concerns, scientists must have systems that expose such considerations early and often to facilitate transparency into data uses.

The broader ecosystem of data, including human and autonomous agents, must consider model creation (by humans and AI agents), deployment of software and algorithms, and human oversight of these processes. Data repositories will hold raw and processed data, software, agents, models, and audit and oversight trails. The manner in which humans retrieve and interact with this ecosystem will be enmeshed with the data, human–AI interfaces, and the management and control plane that cuts across this ecosystem.

Accelerating proliferation of AI methods and applications will require that the data infrastructure adapt to accommodate these transformative technologies. New software pipelines will come together end-to-end, pulling models from diverse sources (Figure 12.3). Advancements in the coming decades will require an increasingly flexible, ever-evolving data and software ecosystem that is changeable, self-tuning and explainable so that human overseers can provide appropriate oversight.

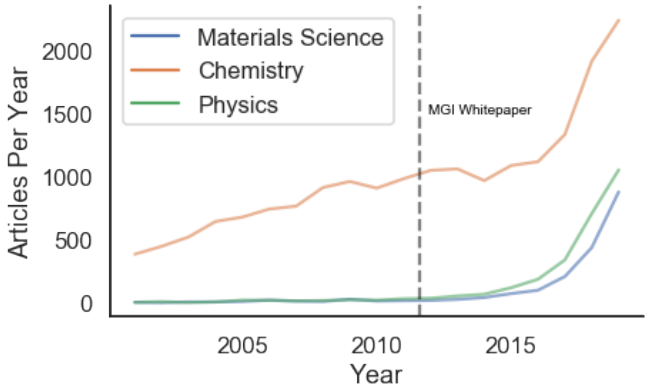


Figure 12.3 *Bibliographic analysis shows rapid growth in the number of papers describing AI/ML methods in different physical science disciplines [11].*

Integrate data and theory to create converged knowledge repositories. Petabytes or even exabytes of data may be essential to progress in AI-driven science, but scientists cannot realistically manipulate all relevant data whenever they ask a question. Instead, as discussed also in Chapter 10, AI Foundations and Open Problems, methods are needed that can summarize large quantities of data in forms suitable for use in subsequent research, and then manipulate both such summaries and explicit symbolic (e.g., mathematical, but also qualitative and natural language–derived) representations of knowledge.

Consider, for example, how Kepler extracted, from decades of observations of planetary motion, compact quantitative relationships among such quantities as orbital period and axis, which coalesced as his Laws of Planetary Motion. Consider also how Newton's Laws of Motion allow for direct calculation of many relationships. Similarly, AI-driven science requires the ability to synthesize new models from data, so that massive data can be consumed effectively by scientists and AI methods, and assimilate symbolic representations of known relationships and physical laws. Technologies, methods, and best practices are needed for synthesizing learned models from scientific data via the use of AI methods, and for working with datasets, learned models, and symbolic knowledge in natural and efficient ways.

The FAIR principles discussed earlier can be repurposed from data to models, yielding the following requirements and research challenges for learned models.

- *Findable*: Innovations are needed to enable discovery of models that meet specific research needs, relating, for example, to domain of applicability, uncertainties, and nature of the source data used to generate them.
- *Accessible*: New approaches to structuring models are needed to allow them to respond to a wide variety of both human- and machine-driven queries. It is important for scientific reproducibility that model outputs can be related to the data elements used to create the model, except when working with sensitive or proprietary data, when models must encode relationships without revealing the specifics of, for example, a single patient's medical record (see Chapter 3, Biology and Life Sciences) or a single manufacturer's drug assay.
- *Interoperable*: As many models are created, it will become important to be able to combine them—to chain together, for example, models of materials properties and manufacturing processes to explore materials that are both nontoxic and manufacturable. Innovations are needed to create models that can be linked in such ways.
- *Reusable*: A model that summarizes a certain physical phenomenon needs to be callable in different contexts, including from within simulations and other computations, and be deliverable to different locations (e.g., supercomputers, edge devices) for different purposes.

Architect new infrastructure to support ubiquitous scientific AI. The infrastructure required to help accumulate and support FAIR principles must be ubiquitously available to support science campaigns from the start, help accelerate discoveries through domain science

advances, and promote the use of AI on the data. Today's examples of autonomous vehicles, Internet and media data, and personal health data demonstrate the needs as well as the constraints on data acquisition, movement, staging, storage, and access. Science domains grapple with different types of data and impose greater constraints than are typically encountered in other settings. For example, scientific data can be several orders of magnitude larger than enterprise data. In addition, data movement needs stretch the limits of current connectivity, despite powerful tools [12].

The traditional infrastructure that supports large volumes of data must evolve to support AI-friendly access. Data volumes and retrieval rates need to scale significantly. AI will require data to train predictive models, observations to infer steps in the process, and control data to modify and optimize the feedback loop of theory to experiment and back to improvements in theory (and our understanding of the world).

Given such a need for the acceleration of AI with data, data access pathways must be scalable in both breadth (distributed) and depth (low latency and high-bandwidth). New search and retrieval techniques that extend beyond the capabilities of our current centralized approaches must be developed to support the ubiquitous reach of AI techniques. Underpinning these new data flows will need to be greatly enhanced computational capabilities, not only centralized but also co-located with data producers and consumers, and configured to support specialized AI workflows. DOE user facilities would serve as ideal test beds for prototype deployment, user testing, and algorithm development.

The underlying data, both feeding the AI and growing as a result of it, must have a provenance and use trail. Data collections and the state and history of associated science campaigns should be easily shareable. Capabilities must explicitly support the dissemination and exchange of scientific data. Such tools and technologies would be central to the needs of scientific reproducibility, providing the capability to validate and trust experiments, improve science campaigns in the future, and dramatically enable the reuse of AI techniques for science. The goal is to develop systems that can *collect data for AI, enhanced by AI—and make those data accessible anywhere in reusable forms*. As the preparation, organization, and use of data for AI becomes streamlined and better understood, the value of the appropriate state of data (raw or reduced) will drive when and how data are retained within its life cycle.

3. Advances in Next Decade

Opportunities and challenges in AI for science are expected to evolve rapidly over the next decade due to three major factors:

Dramatic increases in the volume of available data. The amount of data available is expected to result from improvements in scientific instrumentation, sensors, and computation. For example, the upgraded Advanced Photon Source (APS) at Argonne National Laboratory will produce, from 2023 onwards, up to three orders of magnitude more data than in 2019.

Emergence of autonomous laboratories. Laboratories capable of collecting data about scientific phenomena without human intervention will drive developments that have the potential to transform scientific AI by generating data with greatly increased speed and consistency, but they will also introduce new challenges relating to access and potential algorithmic bias in terms of data collected.

Rapid improvements in AI software and methods. Many industries are working to address massive data and learning problems, such as for autonomous vehicle fleets with tens of millions of vehicles and remote sensing in thousands of small satellites. Work in these and other areas will produce continued

improvements in how data are acquired, organized, and processed, and in the overall data life cycle itself. Developments relating to the integration of symbolic knowledge and data and progress toward artificial general intelligence will produce new datasets, conceptions of how to use data in AI, and methods for manipulating data that may have relevance to scientific AI. DOE science must engage effectively with these developments.

4. Accelerating Development

An early priority for a scientific AI initiative must be to ensure sustained access to the large quantities of high-quality data needed to advance AI-driven science. This means prioritizing work to harness important data flows; to establish the machinery needed to collect, organize, and refine the resulting data; and to gain experience with the use of those data for AI purposes.

Harnessing important data flows means developing and deploying machinery to collect AI-critical data generated by scientific instruments, including the descriptive metadata required for those data to be useful within AI applications. An envisioned program will implement such comprehensive data collection for a dozen different data sources within a year, with this experience guiding expansion to progressively more sources in subsequent years.

Simultaneously, efforts should be launched to develop the technologies and infrastructure required to ingest, organize, annotate, curate, index, and otherwise prepare these data for use in AI applications. Collaborations between Office of Science user facilities and data projects (e.g., NMDC), on the one hand, and ASCR researchers, on the other, can help to accelerate method development. Efforts should also be started to develop AI applications based on the increasingly large quantities of data that will be collected as automated data collection machinery is deployed. Establishing governance of dataset quality, incorporating best practices from diverse communities, and providing intuitive policy and efficient mechanisms for data access and ownership will be vital for data to be available to AI.

A final area of effort should focus on expanding the national workforce and growing the expertise of scientists and other professionals who will prepare, manage, control, deploy, and monitor the data backplane integral to AI. Science campaigns that rely on domain scientists must be structured to include data scientists early in the process. These data scientists and domain experts must work together in an interdisciplinary manner, with explicit cross-training to ensure that the data life cycle contributes to accelerating science goals.

5. Expected Outcomes

This chapter makes the case for the automated creation and use of rich, curated collections of AI-ready scientific data; the integration of large data collections with symbolic representations of scientific knowledge; and new, ubiquitous infrastructure to enable effective use of those data within AI-driven workflows. These developments will transform scientific discovery by enabling better science, faster, and at lower costs; drive virtuous cycles whereby better data produces better AI, and better AI produces better data; and contribute to expanded scientific leadership for DOE and the nation.

6. References

1. Wilkinson, M. D. et al. The FAIR Guiding Principles for Scientific Data Management and Stewardship (*Sci. Dat.* 3, 2016).
2. Blaiszik, B. et al., A Data Ecosystem to Support Machine Learning in Materials Science (*MRS Commun.*, 2019).
3. Himanen, L., Geurts, A., Foster, A. S., Rinke, P. Data-Driven Materials Science: Status, Challenges, and Perspectives (*Adv. Sci.*, 2019).

4. Arkin, A. P., et al. KBase: The United States Department of Energy Systems Biology Knowledgebase (*Nat. Biotechnol.* 36, 7, 2018).
5. Williams, D. N., et al. The Earth System Grid: Enabling Access to Multimodel Climate Simulation Data (*Bull. Am. Meteorol. Soc.* 90, 2, 195-206, 2009).
6. Stokes, G. M., & Schwartz, S. The Atmospheric Radiation Measurement Program (*Bull. Am. Meteorol. Soc.* 75, 7, 1201–1222, 1994).
7. Weber, G. H., Ophus, C., & Ramakrishnan, L. Automated Labeling of Electron Microscopy Images Using Deep Learning (*Proc. IEEE/ACM Mach. Learn. in HPC Environ.*, 26–36, 2018).
8. Biven, L., Office of Science Data for AI Roundtable: Presentation to ASCAC (http://bit.ly/2QWyTbr, 2019).
9. Aspuru-Guzik, A., & Persson, K. Materials Acceleration Platform: Accelerating Advanced Energy Materials Discovery by Integrating High-Throughput Methods and Artificial Intelligence (http://nrs.harvard.edu/urn-3:HUL.InstRepos:35164974, 2018).
10. Carbonell, P. Radivojevic, T., & García Martín, H. Opportunities at the Intersection of Synthetic Biology, Machine Learning, and Automation (*ACS Synth. Biol.* 8, 1474–1477, 2019).
11. Blaiszik, B., Charting ML Publications in Science (https://github.com/blaiszik/ml_publication_charts, 2019).
12. Chard, K., et al. The Modern Research Data Portal: A Design Pattern for Networked, Data-Intensive Science (*Peer J. Comput. Sci.* 4 e144, 2018).

13. Hardware Architectures

AI is a powerful force driving the design of computer architectures [1]. Impacts include both an explosion of new start-ups and hardware designs, and rapid evolutionary change in all platforms—CPUs, GPUs, and even mobile phones. The recent 2019 AI Hardware Summit in Mountain View, California, included nearly 40 companies; enterprise, semiconductor manufacturers, and AI hardware start-ups from around the globe presented their AI architectures and systems.

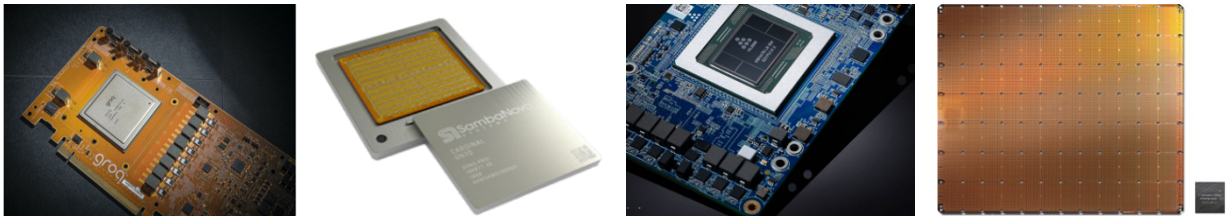
Although these studies and investments are impressive, most of these activities focus on consumer or enterprise areas such as autonomous driving, social networks, finance, and virtual reality [2]. The key problems for these areas are image and video analysis, language translation, and autonomous driving. All of these have data characteristics, real-time requirements, and deployable resource targets that are vastly different from the DOE mission. Specifically, these commercial AI areas have massive numbers of small, labeled data items (e.g., pictures) from which to generate their models. DOE mission areas include areas of computational science with HPC and experimental data, where the dataset can be drastically different: hundreds of simulations or experiments with dozens of dimensions, rather than millions of photos.

For this reason, it is recommended that DOE create a focused strategy to shape AI hardware to serve its science mission. Key to success is a strategy that leverages community and industry investments in technology and scalable (see Chapter 16, Facilities Integration and AI Ecosystem), intermediate, and edge systems (e.g., field instruments, see Chapter 15, AI at the Edge) for AI.

1. State of the Art

DOE user facilities will continue to see increasing data volumes and rates from large experimental facilities such as light sources, nanoscience centers, and advanced computing facilities. As detailed by science domain teams (elsewhere in this document), effective collection and analysis of these data will be enhanced by adopting AI techniques, which will often be deployed using specialized AI accelerators to increase performance and energy efficiency (Figure 13.1). Applying AI techniques to process these data streams requires data management capabilities that can reach from the instrument at the edge to the data center. Without carefully integrated, orchestrated, and managed data infrastructure, these AI systems will not be productive for science. Moreover, these scenarios will introduce additional complexities of heterogeneous hardware (e.g., x86 multicore, GPUs, and specialized hardware like TPUs) [4] and associated programming systems (e.g., MPI and TensorFlow).

Motivated by early results, algorithms and computer architectures for AI are quickly evolving and growing more diverse. These architectures include specialized devices at different scales for five use cases [3]: (1) AI research and development (requiring maximum flexibility for experimentation); (2) offline training of AI models in production; (3) inference on servers; (4) inference at the edge; and (5) online learning on servers and the edge. As other successful AI technologies emerge, such as graph-based ML, the computational challenges and deployment techniques will evolve naturally.



(a) (b) (c) (d)

Figure 13.1 *Examples of AI accelerators from[1] (a) Groq, (b) SambaNova, (c) Habana, and (d) Cerebras Systems.[2]*

At one extreme, systems with thousands of specialized architectures (e.g., NVIDIA Volta and AMD MI60 GPUs, FPGAs from Intel and Xilinx, Google TPUs [4], SambaNova, Groq, Cerebras) are required to train AI models from immense datasets. For example, Google's TPU pod has 2048 TPUs and 32 terabytes of memory and is used for AI model training; its specialized tensor processors provide 100,000 tera-ops for AI training and inference. In addition, they are coupled directly to Google's cloud, a massive data infrastructure (>100 petabytes). The progress of the Google TPU in its use for Alpha Go series of matches demonstrates that codesign—the refinement of hardware, software, and datasets for solving a specific goal—provides major benefits to performance, power, and quality [7].

At the other end of the spectrum, edge devices must often be capable of low latency inference at very low power. Industry has invested heavily in a variety of edge computing devices for AI including tensor calculation accelerators (e.g., ARM Pelion, NVIDIA T4, Google's Edge TPU, and Intel's Movidius) and neuromorphic devices (e.g., IBM's TrueNorth and Intel's Loihi). Experts expect dramatic improvements in the compute capability and energy efficiency of these devices over the next decade as they are further refined. For example, NVIDIA recently released its Jetson AGX Xavier platform, which operates at less than 30W and is meant for deploying advanced AI and computer vision algorithms at the edge using many specialized devices such as hardware accelerators (i.e., DLAs) for fixed-function convolutional neural networks (CNNs) inference. Another example is Tesla's FSD Chip, which can deliver 72 tera-ops ($72x10^{12}$ operations per second) at 72 watts and support capabilities that can respond in 10 milliseconds (driving speed response) with high reliability.

In contrast, DOE's applications can require responses 100,000x faster—100 nanoseconds for real-time experiment optimization in electron microscopy or APS experiments where the samples degrade rapidly under high-energy illumination (see Chapter 14, AI for Imaging).

In terms of software, currently, many consumer applications of AI use software frameworks like Tensorflow, PyTorch, MXNet, Torch, or Caffe2 that hide much of the complexity of the underlying hardware. As mentioned earlier, these frameworks have been developed for video, image, and speech recognition as well as language translation and natural language processing, but they remain in their infancy for processing scientific data. Furthermore, software integration of this AI ecosystem (e.g., PyTorch) with the HPC ecosystem (e.g., MPI and OpenACC) will be nontrivial; significant challenges remain in coupling and potentially unifying these software ecosystems for productivity and efficiency (see Chapter 11, Software Environments and Software Research).

2. Major (Grand) Challenges

Given this spectrum of architectures and their fast pace of change, DOE will need to be actively engaged with the communities of

[1] Permission to use each of the pictures was granted by each of the respective companies.

[2] The Cerebras Wafer Scale Engine is 46,222 mm2; by comparison the largest GPU is 815 mm2.

applications, data management, software, and broader architectures to have timely impact. More specifically, it is recommended that DOE plan to co-design architectures and develop software for a range of heterogeneous systems that span from the edge to the HPC data center. DOE should advocate solutions for priority requirements for AI in science. If appropriate solutions do not emerge from industry, DOE should pursue them internally, leveraging the new communities in open source hardware and through partnerships with other government agencies.

Along these lines, the workshop participants identified several challenges.

Create predictive architecture design tools to enable rapid evolution of AI accelerators for science. Both AI and the use of AI to augment traditional DOE scientific computing are in their infancy. As such, AI workloads are rapidly changing in nearly every dimension: network structure, model paradigm, numerical precision, training approach, training dataset sizes, data types, and batch size (i.e., working set). This transformation is further intensified by the rapid expansion of application scenarios and AI software systems. PyTorch is now the most popular framework used by researchers at the NIPS conference, but it did not exist 5 years ago!

Not only are the AI software artifacts themselves changing, but the process of developing AI software is different from methods used for traditional software. Most noticeably, AI software development focuses much more on the curation, labeling, and preprocessing of training datasets than writing code. Deployment may also be different; AI will use new end-to-end workflows and specialized hardware as they become available. In addition, as AI models are integrated into existing systems, we will need interfaces for embedding those AI components in the traditional scientific applications seamlessly while hiding specialized hardware intelligently. In this regard, holistic design of AI technologies from server to the edge will be paramount. DOE will need predictive methods and tools to frequently characterize, model, and simulate the AI workflows and algorithms to co-design and procure the appropriate architectures for these mixed HPC simulation and AI workloads. Additionally, these combined HPC and AI systems should also be instrumented to provide rich telemetry data that will be critical for this purpose. Hardware should be designed with this requirement in mind.

In fact, there are significant opportunities to co-design heterogeneous compute nodes that leverage the commodity system-on-chip (SoC) ecosystem. Likewise, another key aspect is the development of the memory subsystem to support this heterogeneous compute node and the co-design of the required memory interface controller. Once this heterogeneous SoC processor is designed, system interconnection network fabrics that build on the momentum from prior DOE investments will be needed to create a postexascale, leadership-class, large-scale heterogeneous system architecture.

Create integrated AI workflows and use them to evaluate emerging AI architectures from the edge SoCs to HPC data centers. Given the broad spectrum of efficient specialized AI architectures expected, DOE scientists will need to run workflows using a spectrum of accelerators (including AI). This “extreme heterogeneity” [5] is challenging, requiring scientists to cobble disparate programming systems (e.g., MPI, CUDA, OpenMP), storage systems, and data formats to run simulations on HPC architectures. With the emergence of AI as a primary technique, contemporary AI frameworks (e.g., TensorFlow, PyTorch) will also need to be integrated. Unified frameworks are needed. This challenge will only get more complex in the coming decade as architectures and relevant programming models become specialized. These integrated workflows are a realistic context in which to evaluate the

real impact of AI architectures. The fact that specialization is successful in the AI market is an indication that hardware specialization as a general strategy is logical and could be employed for other high-value scientific applications.

It is recommended that SoC hardware ecosystems be leveraged by the DOE to co-design flexible, heterogeneous computing systems that better integrate AI elements with both scientific hardware for HPC and edge computing for DOE experimental facilities. DOE science domain teams can complement natural trajectories of vendor product plans that fail to meet DOE mission needs. These teams can co-design node and system architecture concepts that specifically address combined HPC and AI workloads to meet DOE mission needs. In addition, DOE discoveries in materials research may directly lend themselves to advanced AI computation (e.g., [9]), and they should be pursued directly.

Furthermore, future AI architectures may provide new opportunities for algorithms in computational science applications. For example, AI support for sparse neural networks can be repurposed for high- performance sparse matrix computation in a conjugate gradient solver. DOE will need to actively explore these opportunities as the new hardware emerges.

Meet the rapidly growing demand for memory, storage, and I/O capabilities of the emerging requirements of AI-enabled science. Current HPC memory and storage systems are architected for traditional HPC simulation-only workloads with relatively small inputs and large outputs, where the access patterns are predictable, contiguous, block-based operations.

AI training workloads, in contrast, must read large datasets (i.e., petabytes) repeatedly and perhaps noncontiguously for training. AI models will need to be stored and dispatched to inference engines, which may appear as small, frequent, random operations.

On the server side, storage systems, such as those that support Lustre and burst buffers, are not designed for and often perform poorly for these read-heavy, random access workloads. The new designs need to include intelligent workflow management systems that can stage data appropriately using additional levels of storage that can facilitate high-IOPs. Likewise, node-local memory hierarchies are relatively small when compared to scientific datasets that will be necessary for training.

At the edge, energy efficiency and performance of the memory system will be critical. Edge devices will need to perform inference concurrently with other tasks; memory capacities will need to increase to support these tasks. This realization has led to pursuit of alternative memory technologies including NAND flash and 3D Xpoint (e.g., Intel Optane), because they offer superior energy efficiency and density. The precise architecture and software system for using these new memories in AI systems remains an open question, given the change in AI architectures and applications.

Enable the incorporation of explicit science domain knowledge into AI systems and hardware to improve robustness and capabilities. AI training typically requires huge quantities of input data, and system behavior may be fragile if it is subjected to stimuli outside of the original training coverage. Many industry applications have massive datasets (e.g., composed of millions of hours of 4K video or billions of photos) that can be used for training, whereas simulation output and scientific data from experiments are much more expensive to produce and often have many more dimensions, rendering them untrainable due to the “sparsity” of the training data. Current AI systems may become fragile when encountering novel situations that lie outside of their initial training dataset, and the AI systems cannot guarantee that answers

satisfy any explicit constraints (e.g., in physics, AI-inferred results must adhere to the law of conservation of energy). For applications that have consequences to human life, such as autonomous vehicles, adding these "instincts" about physics and causality are an urgent priority for industry AI-hardware investments that will enable systems to manage novel situations and to be trained with less data (e.g., overcoming the challenge of "sparsity") [6].

Industry and academia are in the early stages of developing approaches that instill such "instinct" or "physics knowledge" for future AI-hardware offerings, and will also require deep changes in the underlying architectures. But these efforts are far from sufficient to meet DOE mission needs of real-time edge/sensor applications performance. DOE has an opportunity to partner with industry early to drive the generalization and increased capability (low latency) solutions to suit DOE science applications.

3. Advances in the Next Decade

Industrial investment by large-scale cloud companies as well as AI hardware start-ups will continue to drive performance and energy efficiency at scale and at the edge for commercial applications such as image/face recognition, natural language, logistics, voice assistants, and autonomous vehicles. These commercial drivers will infuse AI capabilities broadly, in the scale of data, complexity of function, and robustness that can be achieved. Within the next 10 years, we expect to see the following:

- Introduction of novel AI algorithms, as they are changing quickly and it is difficult to predict popular algorithms for the next decade. Five years ago, LSTMs were new, ResNets were not in use, and transformer networks had not yet been invented.
- Steady increase in the size of largest AI models trainable as well as improvements in training algorithms that reduce the order of growth in training cost per weight. If the largest model training costs continue their current growth rate of 10x/year, economic and environmental consequences will ultimately be the practical limits.
- Steady reduction and plateau in inference latency and cost to commercially important thresholds (i.e., ~5 milliseconds for human and automobile response times).
- Integration of AI acceleration hardware into all mobile/IoT, and server devices.

These advances will be enhanced by the numerous electronics technology initiatives underway such as the IEEE Rebooting Computing, DARPA Electronics Resurgence Initiative, and SRC's activities like JUMP.

4. Accelerating Development

The AI hardware industry is growing by leaps and bounds. It's led by the hyperscale cloud providers (e.g., Google and Microsoft), but there is opportunity to shape the emerging hardware to broader utility for science. The key is to identify leverage points where DOE science applications will benefit, and industry can benefit from features with the generality to address broader markets. Understanding and tracking DOE's growing AI workloads will enable DOE to provide incisive and actionable input to shape future AI architectures. Identification and use of leverage points will drive the creation of new architectures and systems (both software and hardware) that build on broader industry developments to meet DOE's unique needs for sparse learning and the support of scientific discovery. The areas of highest leverage are as follows.

Create new co-design capabilities in DOE to inform strategic action on integrated AI and HPC systems. Computing hardware architectures are evolving in a disruptive fashion, with important innovation coming from small start-ups, large vendors, and cloud service providers. This business ecosystem transformation means that DOE cannot engage

in the traditional fashion of long-term projects with a few large, known players as in existing Pathforward programs. Rather, DOE must invest and create a much larger internal architectural research and development capability. These must then be used to continually assess the landscape, identify important new breakthroughs and partnerships, and accelerate them rapidly into new large-scale system capabilities and DOE facilities to support rapid, leading-edge exploitation of AI across DOE. In some cases, DOE can exploit new hardware to its advantage (e.g., using mixed-precision algorithms with low-precision hardware), while in other cases DOE can work with industry to provide specific new capabilities.

Support AI for HPC and scientific experiments on the edge. Contemporary HPC architectures are designed to support a traditional simulation-only paradigm, where the amount of input data is relatively small when compared to the output, and where the output is not read frequently. Storage systems and memory hierarchies must be redesigned to accommodate this workload change. Moreover, with the addition of AI at the edge to the DOE portfolio, the model for computing within DOE may need to evolve to where specialized AI hardware cooperates with traditional HPC systems to train models before distributing them to low-power inference engines at the edge (see Chapter 15, AI at the Edge).

Drive development of AI systems and hardware that combines explicit knowledge with learned function. A distinctive requirement for AI in DOE's science mission is the need to fuse explicit knowledge with learned function, which is often the goal of ML. While useful in some commercial applications, the purest and strongest form of this fused capability is essential for scientific exploration, and more importantly, the creation of scientifically sound modeling and exploration computations that are the likely foundation for future computational science. DOE should establish a series of specific science-based challenges to drive and shape new AI technologies that fuse explicit knowledge and learned function (see Chapter 10, AI Foundations and Open Problems).

Lead on ultra-low latency and low-power inference for scientific experiment control in experimental facilities. DOE facilities and experiments are multimillion- (and sometimes billion-) dollar investments that literally push forward the frontiers of knowledge in materials, physics, biology, and other areas of fundamental science. AI-based real-time intelligent control of these facilities cannot only enable more complex, intelligent experiments and more efficient operation, but can directly accelerate the advance of scientific discovery. DOE should charter cross-disciplinary centers and focus on low-latency inference challenges for scientific experiment control, a critical capability within national laboratories.

Translate DOE fundamental materials–device discoveries into new post-CMOS AI devices. There is an opportunity to bring more of the fundamental materials science advances from DOE to augment industry roadmaps through fundamentally new approaches to neuro-inspired AI architecture. This can lead to a new path for exponential growth in AI computational performance by overcoming the overheads of conventional digital hardware, which is the predominant approach for today's AI systems, and software design, addressing the societal challenge of rapidly growing negative environmental impacts of DNN-based AI. This has a strong connection with the Basic Research Needs for Microelectronics [10] activity and collaboration opportunities that span the entire Office of Science (see Chapter 1, Chemistry, Materials, and Nanoscience).

Create and adopt new operational and life cycle models in large-scale DOE computing facilities that support sustainable AI computing. At 10x annual model size increases, the training of large AI models already matches the lifetime carbon emissions

of five gas-powered automobiles [7]. The rise of renewable energy generation creates an opportunity to sustain AI and HPC's growing computing and energy appetite. The DOE can lead in convening its own cloud and academic centers with technology leaders to study and prove new operational models. These models can enable high levels of renewable energy in power grids while sustaining high-capability computing. Further benefits can arise from new life cycle models that shift compute resources from high-cost to low-cost (and low-carbon power) locations that increase the lifetime of hardware, reducing cost and e-waste. To sustain the exponential growth in computing capability at the heart of its scientific missions, the DOE should lead the community in creating and deploying such practices to contain or even reduce its environmental impact for AI and HPC computing (see Chapter 16, Facilities Integration and AI Ecosystem).

5. Expected Outcomes

Industry will continue its dramatic pace of advancement over the next decade, but those advances are focused on goals that will not lead to meeting the requirements of DOE computational science and experimental data applications. In particular, the AI use cases for scientific applications will differ significantly, requiring extreme data rates, low-latency response, and extensive exploitation of explicit knowledge. Second, the rapid growth of AI training costs will create sustainability challenges to the growing AI computing burdens, forcing new approaches. Scientific applications and experiments are likely to have fewer samples available and require more data integration for training. By working together with industry to augment their hardware platform offerings, we will be able to meet these critical needs for the future of AI for HPC, for automating control systems at DOE user facilities, and for creating intelligent sensors for the future of experimental science.

6. References

1. Chien, A. Computer Architecture: Disruption from Above, *Commun. ACM* **61**, 9, 2018.
2. Wu, C., et al., Machine Learning at Facebook: Understanding Inference at the Edge, Proceedings of the 2019 IEEE International Symposium on High Performance Computer Architecture (HPCA), 331–44 (https://doi.org/10.1109/HPCA.2019.00048, 2019).
3. LeCun, Y. Deep Learning Hardware: Past, Present, and Future, Proceedings of the 2019 IEEE International Solid-State Circuits Conference (ISSCC), 12–19 (https://doi.org/10.1109/ISSCC.2019.8662396, 2019).
4. Jouppi, N. P., et al., In-Datacenter Performance Analysis of a Tensor Processing Unit, SIGARCH Comput. Archit. News **45,** 2,1–12 (https://doi.org/10.1145/3140659.3080246, 2017).
5. Vetter, J. S., et al., Extreme Heterogeneity 2018 – Productive Computational Science in the Era of Extreme Heterogeneity: Report for DOE ASCR Workshop on Extreme Heterogeneity, USDOE Office of Science (https://www.osti.gov/servlets/purl/1473756, https://doi.org/10.2172/1473756, 2018).
6. AAAS Science Magazine, How Researchers are Teaching AI to Learn Like a Child, May 24, 2018.
7. Strubell, E., Ganesh, A., and McCallum, A. Energy and Policy Considerations for Deep Learning in NLP. (https://arxiv.org/abs/1906.02243)
8. Silver, D., et al., Mastering the game of Go without human knowledge, *Nature* **550**, 354, (https://doi.org/10.1038/nature24270, 2017).

9. Torrejon, J., et al., Neuromorphic computing with nanoscale spintronic oscillators, *Nature* **547**, 428, (https://doi.org/10.1038/nature23011, 2017).
10. Basic Research Needs for Microelectronics (Brochure), USDOE Office of Science (https://www.osti.gov/servlets/purl/15457, 72, 2018).

14. AI for Imaging

The DOE-supported x-ray, neutron, electron beam, and nanoscale science research centers are major experimental facilities providing access to world-class imaging and physical characterization capabilities to more than 14,000 visiting scientists and engineers annually. The advanced tools and instruments at these facilities probe complex materials and processes across the physical, materials, environmental, and life sciences, including those underpinning energy technologies and advanced manufacturing. Research done at these user facilities provides unique insights that help shape the future and ensure the economic competitiveness of the United States. Planned developments at these facilities over the next decade promise to produce vastly larger and more complex datasets much more quickly than today, making the automation of facility and instrument operations and data collection and reduction imperative.

AI will be essential for ensuring continued technological progress and maintaining America's leadership position in all branches of science. The application of AI at DOE user facilities will ultimately allow end-to-end control of the scientific endeavor at scale, improving stability in experimental equipment and processes and yielding superior results. Using AI technologies to augment and expand existing data analysis techniques will allow scientists to process data more efficiently and effectively than ever before. AI technologies could one day make fully autonomous decisions on measurement strategies, reducing experimentalists' time while simultaneously enabling efficient exploration of complex experimental and sample configurations.

1. State of the Art

Modern research laboratories present challenges for control systems that are capable of meeting evolving requirements. For example, consider the particle accelerators driving large-scale research facilities, which consist of many interconnected subsystems of magnets; mechanical, vacuum, and cooling equipment; power supplies; and other components. These accelerators have many thousands of control points, making their operation a complex optimization problem. This is particularly true for the electron accelerators at DOE's synchrotron light sources, which require a very high level of stability. The operation of these accelerators has benefited from AI/ML–based solutions but remains extremely difficult due to the lack of *a priori* models for reliable and safe control. In the absence of such models, learning models based on raw data and other AI/ML-based solutions have been explored, with promising results, beginning with the demonstration of artificial neural network–assisted control of ion sources in the early 1990s. AI/ML optimization methods such as genetic algorithms and particle swarm optimization have been successfully applied for several years to improve various aspects of facility operation, including electron beam lifetime, transverse coupling, and injection efficiency. Current efforts are focused on simulation of the data generated by accelerator physics models to optimize the performance of next-generation machine systems under development [1]. However, none of these advancements have become an integral part of today's accelerator control systems. This is due to limitations in the available data as well as software and hardware infrastructure and the reluctance of communities to use AI as a general purpose tool.

The high data generation rates of modern detectors provide additional challenges for data processing and management at these facilities. As an example, advances in neutron detector technology has enabled high-resolution, 3D tomographic reconstruction of complex, multi-component materials as illustrated in

Figure 14.1. One major challenge is that the increasing data volumes will require autonomous methods for data processing. Several supervised and semisupervised data processing workflows based on different neural network architectures have been proposed for different parts of the data life cycle. For example, the transmission x-ray microscopy instrument at the Advanced Photon Source [2] uses DL for fully automated correlative segmentation of metallic alloys by classifying features in large nano-resolution 3D reconstructed volumes. Similar types of network architecture have also been useful for recognizing known features in data sets and completing tasks such as classifying peaks, working with low-resolution data or assigning theoretical models to reduced datasets [3–5]. Using such an approach can be of great help in interpreting the data, especially when the measured data is complex or when it contains features that are not directly related to the material under study. Integration of AI with these versatile imaging techniques can enable analysis of extremely large data volumes in relatively short time frames while exponentially accelerating tomographic data analysis, possibly opening up novel avenues for performing 4D characterization experiments with finer time steps. More progress is needed and will rely on the generation of curated datasets, robust data processing techniques, and sophisticated data management solutions.

The management of large-scale experimental data will also require smart data reduction techniques. For example, current coherent diffraction imaging experiments can generate data at 3 GB/s, resulting in datasets containing tens of terabytes for a single experiment. The data acquisition process is typically stalled when data generation rates are so high that experimental data can be flushed to nonvolatile memory. This high-volume data acquisition not only extends the end-to-end experimentation time but also limits experiments with time-sensitive phenomena. Advancements in synchrotron light sources, such as the APS and

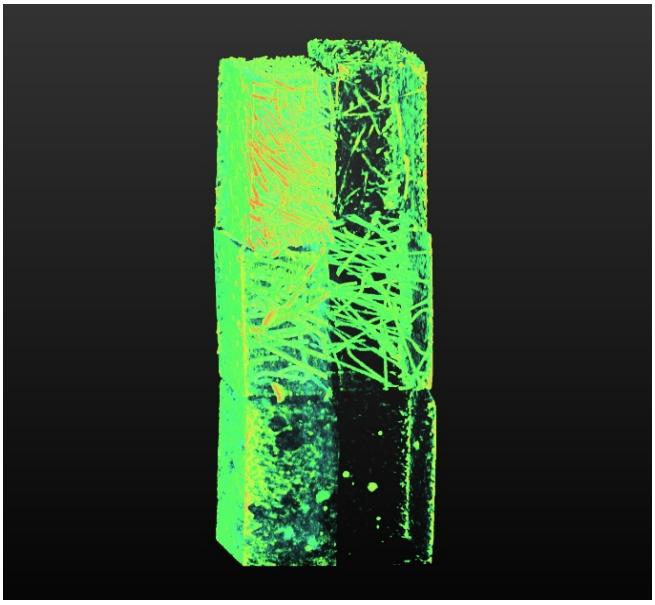


Figure 14.1 *Neutron computed tomography of ultra-high performance concrete (UHPC) is capable of identifying three different phases in the samples: cementitious paste, voids, and H-rich components (reinforced with fibers when present).*

Advanced Light Source (ALS) upgrades, will further complicate these problems as the data generation rates of the detectors will increase by several orders of magnitude for similar imaging modalities. Smart data reduction techniques (e.g., filtering relevant data or point-of-interest data acquisition) will be necessities rather than features with the upcoming light source advancements. Although existing practices work well currently, there is an acute awareness that this model is unsustainable in the long run without the development of additional software and hardware infrastructure and the continuous support of existing activities such as domain-specific AI-based data compression schemes, searchable databases containing both experimental data and metadata, on-the-chip data reduction, and novel algorithms and workflows to improve performance. Some of these tasks related to scientific data management are currently being addressed in the Data Center Pilot, a large collaboration of data and experimental scientists at all U.S. light sources that aims to provide a sustainable road map to the future of data issues at experimental facilities. Similar initiatives are required to develop and maintain

other parts of the large-scale AI ecosystem for research facilities.

2. Major (Grand) Challenges

A new era is dawning in science and engineering, one that promises a revolutionary understanding of complex materials and chemical processes across the entire hierarchy of relevant length and time scales. This understanding demands moving beyond exploration of equilibrium phenomena and beyond models based on idealized materials and systems to create new states and achieve extraordinary new functions [6]. The overarching grand challenge facing scientists at DOE experimental user facilities is to understand, predict, and ultimately design emergent behavior in complex materials and systems. This will require progress in the following areas.

Characterize biological function across length scales. X-ray, neutron, and electron methods generate structural, organizational, and dynamic data across a range of length scales from atomic to mesoscale. Example applications include high-resolution imaging of complex neuronal networks in brains to provide a clearer understanding of how even the smallest changes to the brain play a role in the onset and evolution of neurological diseases, such as Alzheimer's and autism, and perhaps lead to improved treatments or even a cure. Figure 14.2 illustrates the power of a machining learning approach that maintains signal-to-noise ratios with shorter x-ray exposures minimizing damage to radiation sensitive mouse brain. A second example is the characterization of heterogeneous systems such as the rhizosphere, where bacterial communities synergistically interact with the soil and roots of plants, and the observation of the dynamic assembly of functional complexes interacting within living cells [7]. The neutron radiograph in Figure 14.3 demonstrates the role of water dynamics in this region. In both these examples, molecular interactions at the nanometer scale cause emergent behavior at the millimeter scale and ultimately govern the dynamics of complex biological systems and tissues of interest. AI/ML methods will play a critical role in linking multimodal observations across this large, dynamic range of scale and is needed to build a predictive understanding of biological function across time and space.

Figure 14.2 *An image showing individual myelin sheaths, highlighted with different colors, surrounding mouse brain axons revealed by the analysis of experimental nano-CT scans taken at the 32-ID beamline of the APS [11].*

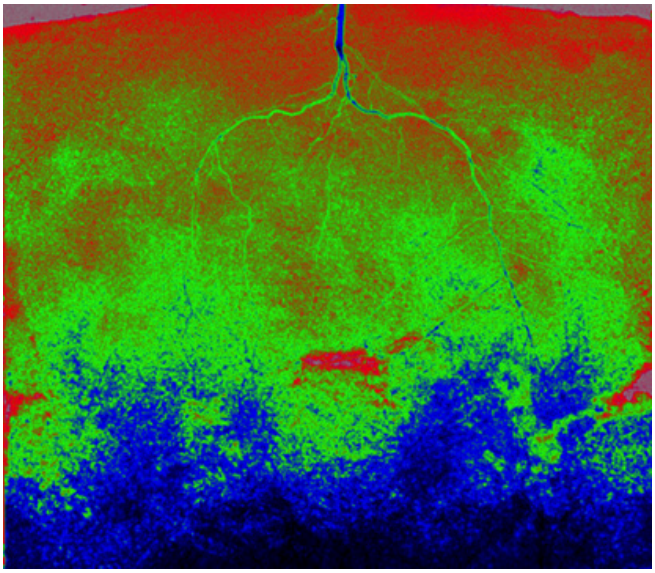


Figure 14.3 *Neutron radiograph of droughted black cottonwood (*Populus trichocarpa*) root system that has been rehydrated to measure the water dynamics in the roots, rhizosphere, and bulk soil regions. Worked performed at HFIR CG-1D Imaging beamline.*

Observe and control nanoscale chemical transformations in macroscopic systems. Devices currently in use or being developed for selective and efficient heterogeneous catalysis, photocatalysis, energy conversion, and energy storage rely heavily on diverse multiscale phenomena, ranging from interfacial electron transfer and ion transport occurring on nanometer and picosecond scales to macroscale batteries that charge in hours and catalytic reactors with turnover times of seconds. Existing and planned instrumentation at DOE user facilities can probe these environments with atomic, chemical, and isotopic contrast spanning a large spatiotemporal range, thereby providing unique fundamental information about these functioning mesoscale chemical devices. These "nanokinetic" operando measurements are essential to optimize complex multiscale chemical and electrochemical devices. The use of AI/ML tools will be essential in interpreting the data by integrating experimental observations with computer modeling to provide multiscale models of complex chemical processes of importance to DOE's mission.

Understand and characterize physical and chemical processes in extreme environments. Understanding materials and processes in extreme environments such as ultrahigh pressure or temperature is of vital importance to the development of fusion and fission materials. Furthermore, such understanding provides unique insights into planetary physics and geosciences. The study of these materials using spectroscopic-, diffraction-, and imaging-based methods is performed throughout the DOE experimental user facility complex and, with the advent of diffraction-limited synchrotron light sources and next-generation free electron lasers, will provide structural characterization modalities of matter in extreme states that are far beyond what is achievable today. The use of AI/ML will be essential in the data analysis of these systems for the reduction of noise, solution of inverse problems, and linking of observations to molecular simulations.

To solve these pressing scientific problems, new tools need to be developed. While AI/ML will play an important role in achieving these objectives, a number of synergistic, basic scientific and engineering developments will also be needed, including high-performance x-ray and neutron optics; sample chamber environments improved electron, x-ray, and neutron detectors; sample handling robotics; and upgrades to computational infrastructure for the efficient movement and storage of data across HPC facilities.

3. Advances in the Next Decade

A number of advances planned for the next decade will necessitate the need for AI/ML tools. There is a need to understand and model the behavior of complex systems across length scales and modalities to transform basic scientific discovery into a set of engineering principles that allow scientists to provide solutions to problems such as the need for plentiful, safe drinking water, safe, efficient alternative energy sources, and therapies to address degenerative diseases. While AI/ML will provide an indispensable set of tools to model these systems, parallel developments in improving neutron and (light) sources for x-ray and electron microscopy and hyperspectral imaging are fueling a revolution in the physical characterization of samples, with an attendant need for AI/ML support. A selected set of developments is summarized below.

The development of diffraction-limited synchrotron light sources. The planned upgrades to Argonne's ALS and APS will yield hard, tender, and soft x-ray sources [8,9] with significantly increased brightness that will allow scientists to explore more complex and disordered samples under controlled and operating conditions with a precision unknown before now—literally approaching theoretical limits.

The development of new hardware for electron microscopy. Next-generation detectors for electron microscopy will greatly

increase data rates, resulting in better prediction of electron strike locations. They will also improve the electron beam–induced sample motion correction that has historically been a resolution-limiting factor. In addition, the development of "phase plates" will increase the contrast between sample and background, revealing particles that traditionally have been too small for electron microscope imaging. These and additional developments in sample delivery will significantly increase the applicability of electron microscopy, as well as the need for data management tools.

The development of ultrafast light sources. The development of ultrafast light sources such as LCLS-II will be transformative for energy science as it will qualitatively change the way in which x-ray scattering, spectroscopy, and imaging can be used. High-repetition-rate machines will enable imaging of natural and artificial systems, spanning multiple decades of time scales and multiple spatial scales. High-repetition-rate sources will enable powerful new ways to capture rare chemical events, characterize fluctuating heterogeneous complexes, and reveal underlying quantum phenomena in matter using nonlinear, multi-dimensional, and coherent x-ray techniques that are only possible with a true x-ray laser.

The development of higher brightness neutron sources. The planned upgrade of ORNL's SNS accelerator to higher power and the addition of a Second Target Station will enable more rapid, time-resolved measurements of transient and out-of-equilibrium phenomena; exploration of matter at extreme conditions, such as magnetic field, temperature, and pressure; and simultaneous measurement across broad ranges of length, energy, and time.

4. Accelerating Development

To accomplish the grand challenges listed above and optimally leverage the hardware advances planned for the next decade, the use of AI-based tools is an absolute necessity. The ability to improve the stability of the instrumentation, perform experiments in an autonomous fashion, and interpret scientific data in a fully automated workflow—and the ability to discover patterns and behaviors across multiple experiments—will greatly accelerate scientific discovery (see Chapter 10, AI Foundations and Open Problems and Chapter 12, Data Life Cycle and Infrastructure). To facilitate this vision, investments in scientific data warehousing and real-time, experimental steering infrastructure need to be made (see Chapter 12, Data Life Cycle and Infrastructure), facilitated by state-of-the art data streaming and edge computing strategies (see Chapter 15, AI at the Edge). At present, there is a lack of tagged or labeled (both raw and processed) scientific data accessible across the DOE landscape, limiting the development and training of AI-based tools that can be deployed in the control and analyses of experiments at user facilities. While facilities are developing AI/ML-based approaches aimed at real-time in-experiment decision making, they are still far from being used routinely. The needed developments are discussed below.

A database that consists of raw detector readings, processed data, and related user proposals, and associated scientific interpretations in the form of standardized data formats and domain-specific metadata languages. Creating enough model or simulated data to provide useful ML training sets will require access to HPC resources. These databases can be built in coordination with user communities, which in turn could be used to train efficient data reduction algorithms, perform data mining operations for the discovery of hidden statistical relations only visible in large datasets, and to build a fully automated "raw data to final model" analysis pipeline. Ideally, facility staff, users, and research communities in the broad sense would aid in a "data-tagging campaign" as part of the execution of their experiment.

A database that consists of metadata, such as scientific instrument responses (e.g., flux and focus) in combination with a record of instrument configurations (e.g., motor positions, neutron chopper phases, and monochromator bending parameters) and measurable instrument and environmental parameters (e.g., ring current, cooling water flow, and temperature readings). This information could be used to build advanced predictive models of accelerators, end stations, and sample delivery systems and to aid in automated alignment and calibration of instruments, stabilizing user operations, predicting and preventing catastrophic failures, and/or reducing the total downtime of the instrument. While it is unclear whether data from different facilities can directly be used to inform models, the ability to find common patterns could provide cross-cutting improvements.

AI-guided real-time experimental steering infrastructure based on curated data and metadata of the sample and instrument state during the experimentation. Gaining transformative insight into dynamic materials processes requires identification, tracking, and quantification of the most relevant volumes within a sample under various conditions of applied stimuli. AI tools have been shown to be capable to provide this type of guidance [10] and should ultimately suggest alternative imaging modalities with which to query the volume of choice. This situation presents a vast measurement parameter space that cannot be exhaustively surveyed, and that is very difficult to navigate when seeking concrete connections between sparse local phenomena, such as dislocation motion and grain boundary stress concentration, and bulk properties as a function of environment, especially in the context of irreversible processes.

5. Expected Outcomes

Given the anticipated pace of development in imaging, scattering, spectroscopy, and associated hardware, there is a dire need to develop data analytics technologies that can aid in making the best choices in experimental design, data reduction, and model generation while reducing the overall cost of data transmission, annotation, and storage.

Early successes in the use of AI/ML tools in experimental facilities indicate that AI/ML will enable the throughput of experiments via autonomous experiments, resulting in the ability to explore larger sample configurational setups in a shorter amount of time, yielding more complete and informative scientific hypotheses. This in turn will result in a reduced cost of discovery, bringing scientific solutions to industry and society at a faster pace and at a reduced cost. AI/ML for the control of equipment will significantly reduce the need for human intervention in tasks that are currently performed by hand, resulting in more stable experimental facilities that produce superior data with a lower rate of human intervention. Similar arguments can be made for the data analysis component, resulting in more scientific opportunities and better use of the in-demand resources available at major DOE user facilities.

6. References

1. Liu, S., Leemann, S. C., Hexemer, A., Marcus, M. A., Melton, C. N., Nishimura, H., Sun, C. 2019. Demonstration of machine learning-based model-independent stabilization of source properties in synchrotron light sources. *PRL*, in press.
2. Kaira, C. S., et al. Automated correlative segmentation of large transmission x-ray microscopy (TXM) tomograms using deep learning. *Mater. Charact.* 142, 203–210 (2018).

3. Pelt, D. M., & Sethian J. A. A mixed-scale dense convolutional neural network for image analysis. *PNAS*, 115 (2) 254–259 (2018).
4. Chang, M.C., et al. Accelerating neutron scattering data collection and experiments using AI deep super-resolution learning (arXiv:1904.08450, 2019).
5. Samarakoon, A. N., et al. Machine learning assisted insight into spin ice $Dy_2Ti_2O_7$ (arXiv:1906.11275, 2019).
6. U.S. Department of Energy. Report from the Basic Energy Sciences Advisory Committee. Challenges at the frontiers of matter and energy: transformative opportunities for discovery science (2015).
7. U.S. Department of Energy. Report from the Biological and Environmental Research Advisory Committee. Grand Challenges for Biological and Environmental Research: Progress and Future Vision; A Report from the Biological and Environmental Research Advisory Committee, DOE/SC–0190, BERAC Subcommittee on Grand Research Challenges for Biological and Environmental Research (science.osti.gov/~/media/ber/berac/pdf/Reports/BERAC-2017-Grand-Challenges-Report.pdf, 2017).
8. The Advanced Photon Source Strategic Plan: Enabling frontier science in the national interest (2018).
9. ALS-U: Solving Scientific Challenges with Coherent Soft X-Rays (2017).
10. Noack, M. M., et al. A Kriging-Based Approach to Autonomous Experimentation with Applications to X-Ray Scattering, *Sci. Rep.* **9**, 11809 (2019).
11. Yang, X., et al. “Low-Dose X-Ray Tomography through a Deep Convolutional Neural Network.” *Sci. Rep.* **8**, 2575 (2018).

This page intentionally blank.

15. AI at the Edge

Many of the use cases outlined in previous chapters—Chapter 4, High Energy Physics, Chapter 14, AI for Imaging, and Chapter 16, Facilities Integration and AI Ecosystem—describe scientific discoveries using large instruments such as the Large Hadron Collider, the Very Large Array, and the IceCube South Pole Neutrino Detector. Likewise, DOE operates many distributed facilities, such as the ARM Climate Research Facility, that operate sensors and instruments across the planet (see also Chapter 2, Earth and Environmental Sciences). For both centralized and distributed facilities, instruments such as these produce vast quantities of data that often cannot be efficiently moved to or stored in a central repository, or they include latency-sensitive control systems that must act promptly on the incoming data. Moving a portion of the data analysis pipeline "to the edge," where the data is generated, allows the required computation to identify the highest value data to be saved and to autonomously respond and control the experiment. The potential benefits of edge computing are widely recognized, and a considerable amount of work to realize and expand upon these benefits in business and science is under way [2].

Advances in AI and ML, both in hardware and software, are among the enablers of edge computing. For example, edge computing enables a self-driving vehicle to make decisions within the vehicle, using AI techniques to interpret data from the vehicle's many cameras and sensors. This is necessary both because of the volume of data (i.e., too large to transmit to central servers) and the real-time requirement for vehicle controls (i.e., answers from remote servers may arrive far too late). Edge computing is possible, even with relatively low-powered computing hardware in the vehicle, because a large body of training data has been processed on high-performance servers (i.e., in the center) into ML models that can be deployed to run in the vehicle (i.e., at the edge).

In the DOE community, a large and growing number of science and engineering projects require edge computing to imbue sensors with real-time adaptive or autonomous capabilities. In addition to the examples mentioned in Chapters 4, 14, and 16, consider the following. There are thousands of environmental monitoring sensors that typically produce longitudinal data with latencies of minutes to weeks between measurement and data availability due to their remote locations and low (or intermittent) capacity network connectivity (see also Chapter 3, Biology and Life Sciences). Edge computing capabilities would enable such instruments to analyze data locally in real time and feed a lower volume of processed information to central computing services for further processing. A radar deployed by the DOE ARM facility in Oklahoma could use ML at the edge to identify important weather phenomena and dynamically steer the instrument for more precise follow-up observations. Such an approach would increase the accuracy and timeliness of tornado warnings, ultimately saving lives. As mentioned in Chapter 8, Smart Energy Infrastructure, monitoring electrical power distribution infrastructure could prevent power failures or predict conditions conducive to wildfires; monitoring subsurface vibrations from oil wells could improve oil production; autonomous soil sampling and analysis devices could improve crop yield; more timely data analysis options would enable large-scale accelerators and light sources to optimize their operations and predict (and prevent) failures.

DOE is in a unique position to address these challenges because it supports many of the research facilities requiring edge computing, either in the near term to better operate existing instruments or in the longer term to

develop more intelligent instruments. Furthermore, DOE supports an extensive community of scientists working on technologies including high-performance sensors, detectors, ML techniques, new computing architectures, and other critical facets of AI technologies required to tackle the grand challenges of today and tomorrow. As illustrated in Figure 15.1, DOE has the computing resources necessary to develop increasingly sophisticated models to run at the edge, sensor capabilities to support its many facilities and instruments, and edge computing research platforms to demonstrate the potential for enabling new science. Additional details on algorithm and software environment research are given in Chapter 11, Software Environments and Software Research, and Chapter 10, AI Foundations and Open Problems.

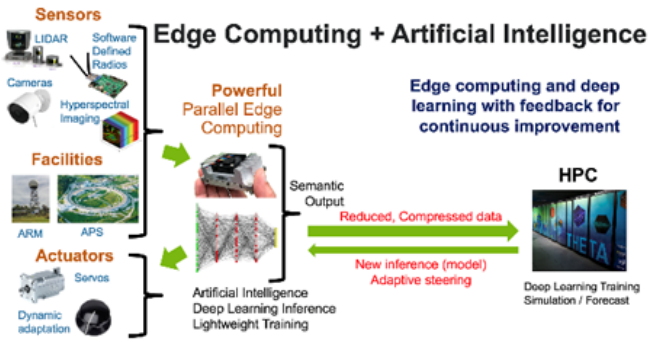


Figure 15.1 *Illustration of edge computing, from the sensor or instrument to the high-performance computer or cloud and back. [Presented in 2018 at the DOE Advanced Scientific Computing Advisory Committee (https://science.osti.gov/-/media/ascr/ascac/pdf/meetings/201809/ASCAC-EdgeAI-Beckman.pdf.)*

1. State of the Art

Experimental facilities such as those operated at national laboratories have been generating large amounts of data traditionally provided to users via removable storage media. Upgrades and improvements to these facilities in recent years have increased the data volume to the point where such methods are impractical. As a result, they have joined the unique facilities noted earlier in implementing edge computing data management and analysis services local to their instruments. More importantly, these edge computing capabilities allow experiments to be adaptively steered. For example, Figure 15.2 shows a series of images illustrating an automation system under development where an electron microscope is first used in a low-resolution mode, while some AI algorithms (running locally with the instrument) identify regions on the sample with features of interest. The electron microscope is then directed to scan the selected regions in a higher resolution mode.

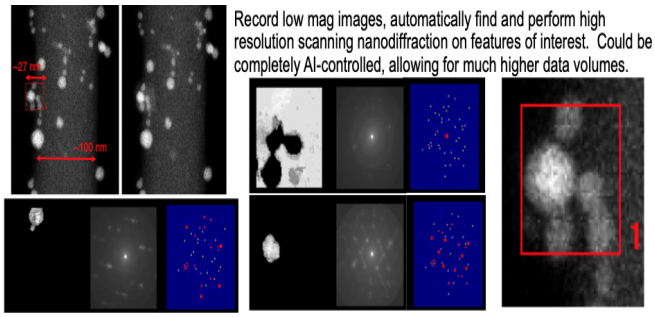


Figure 15.2 *Edge computing can enable instrument steering using AI at the edge to identify features of interest [3].*

In addition to the need for edge computing with centralized instruments, such as electron microscopes and light sources mentioned above, there are applications that require input from networks of sensors. For example, DOE funds a number of environmental monitoring projects where sensors are distributed, often in remote locations with limited networking connectivity. In these cases, edge computing is required both for data compression and for adaptive sensing. Similar to the previous example with electron microscopy, high-bandwidth measurements may only be necessary during events of interest, and edge computing could change the sensor's sampling rates in the same fashion.

Rapid improvement of low-cost sensors is creating opportunities for unprecedented measurement capabilities, all of which require edge computation. But in the absence of general-purpose edge computing platforms, most teams are creating ad hoc solutions. For example, in a large experiment involving monitoring of black carbon in the urban environment [4], scientists at LBNL had to limit

the sampling rate of the sensors so that the amount of data could be easily handled by the available network. In other cases, such as Argonne's Waggle project [5], multiple science groups pooled expertise and resources to build a shared, general-purpose edge computing platform. While the resulting devices were more expensive than traditional sensor devices without edge computing, the cost was shared by multiple experiment teams in large-scale projects such as Chicago's Array of Things [6], an experimental instrument with dozens of sensors. The platform also supports industry collaboration on the use of edge computing to create new measurement and fault prediction capabilities for the national electric power grid [7], where more precise monitoring and analysis of electricity generation and loads could enable AI-based models to forecast catastrophic power failures.

With additional advances in sensing technology and AI capabilities such as Google Edge TPU [8], Intel Movidius [9], and IBM TrueNorth [10], the use of AI at the edge will continue to grow. Integrating these advances into DOE mission-critical applications could dramatically improve scientific productivity.

2. Major (Grand) Challenges

While industry is certainly interested in AI at the edge, its focus is largely on delivering AI products to end users (e.g., cell phones and autonomous vehicles). DOE will leverage the technologies developed for industrial applications; however, it will need to address some unique challenges for scientific applications. The following sections present the grand challenges in the scientific arena that will motivate the computer science and applied mathematics work necessary for supporting AI at the edge.

Improve scientific productivity with high-speed data through AI at the edge. One unique challenge with scientific applications is that they often involve very-high-speed sensors. For example, devices with data rates of 100 TB per second are currently being tested in cryogenic electron microscopy. A number of other light sources and electron microscopes are expected to return similarly high data rates in the future. AI at the edge could be an effective way to process such high-velocity data streams.

Another example is distributed acoustic sensing (DAS), which uses fiber-optic cable to monitor seismic motion. It is much cheaper to deploy than traditional seismometers and captures the motion along the full length of the cable. It has the potential to revolutionize many subsurface applications but has to quickly analyze a large volume of data (i.e., terabytes per day). Fortunately, the common data analysis procedure illustrated in Figure 15.3 only requires the raw sensor data in the first step; therefore, DAS is amenable to edge computing to generate interferometry. Because interferometry is much smaller in data volume than the raw data, it can be much more easily shipped to a central location for further processing.

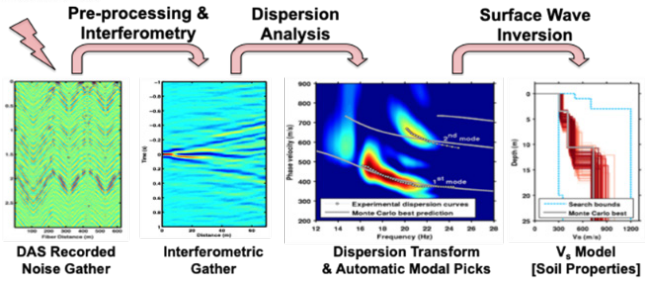


Figure 15.3 *Steps in data analysis for distributed acoustic sensing.*

There are many high-speed data streams that could be similarly processed, which makes processing high-speed data streams the first challenge for AI at the edge.

In addition to data velocity, AI at the edge will also have to deal with data quality issues arising from malfunctioning sensors while working under constraints of memory, storage, and electric power.

Enhance scientific discovery through integration of multiple data sources. Scientific applications not only analyze data from each high-speed sensor separately, they also need effective integration of up-to-date information from multiple, often heterogeneous, sensors. AI at the edge that can leverage heterogeneous data sources will enable real-time optimization of these scientific applications and new scientific discoveries. The following example from agricultural land use, which has sweeping consequences for carbon sequestration, water resources, chemical pollution, and crop yield, illustrates the complexities involved [11].

The AR1K smart farming consortium [12] envisions integrating data from satellites, sensor-equipped drones, ground stations, and embedded sensors to understand dynamics such as the role of soil conditions and microbiomes in sequestering carbon. Analysis from DNA sampling and sensor data (e.g., soil nutrients and composition) and drone-mounted multispectral cameras aims to improve crop yield while reducing the use of chemicals and fertilizers. Today, many measurements rely on manual sampling and operation, severely limiting the scope and scale of measurement and analysis. AI at the edge would perform data analysis and reduction *in situ* and support a transition to autonomous, robotic soil and plant sampling and sensing devices. Furthermore, it would do so at low cost and small physical scale to enable deployments of thousands of units. This level of scale and automation will be essential for understanding and optimizing the nation's agricultural resources and addressing the growing impact of agricultural land use, ranging from chemical runoff to inefficient use of water.

Many other DOE-supported environmental applications such as the Next-Generation Ecosystem Experiments (Tropics and Arctic) have similar needs to integrate information from many different sensors. Other application areas such as cosmology are also starting to integrate multiple sensors, including optical telescopes, microwave telescopes, and gravitational sensors, to capture and analyze important events.

Such large-scale data integration will likely primarily be used in science in the next decade and will require funding from agencies such as DOE, making it the second challenge for AI at the edge.

Enable smart scientific infrastructures through AI at the edge. The challenges discussed so far involve discrete analysis operations at the edge. To manage large distributed systems, it will be necessary not only to analyze data from multiple sources, but also to make coordinated decisions to direct distributed actions. The DOE ESnet [13] exemplifies aspects of this challenge.

DOE has invested in large, national-scale facilities and cyberinfrastructures. The ESnet that connects these DOE facilities to one another and to the world is central to efficient operations of these facilities. Over its 40-year history, the volume of data flowing through ESnet has consistently doubled every 17 months. Over the next decade, the projected exponential growth of "wearables" and IoT devices will bring new challenges with respect to scale, emergent properties, and cybersecurity. Consequently, cyber-infrastructure operation—from local instrument to facility to laboratory scale to ESnet—will require embedded edge AI to make intelligent decisions and coordinate actions across the globe. To facilitate this, DOE is planning to implement advanced telemetry on ESnet networking equipment and to use AI at the edge to digest and process telemetry measurements to initially recommend and ultimately autonomously execute the majority of the networking operations. These tasks include routing of traffic, fault detection, isolation, and remedy, as well as identifying and addressing cybersecurity threats. This autonomous self-managing, self-tuning, and self-healing scientific cyberinfrastructure, like the mobility ecosystem, will rely on edge AI for

such optimizations. It will also rely on edge AI to forecast, detect, and resolve system-level interactions leading to unforeseen global system behavior, while aiming to optimize for system-wide performance goals.

Other large DOE resources may require similar smart technology to integrate different types of computing resources, sensors, and actuators. In particular, we anticipate a computer center would include a collection of computing resources of different types and sizes (e.g., traditional HPC systems augmented with specialized AI processors and quantum processors). Such a computing facility might be closely linked to nearby experimental facilities, such as light sources and particle accelerators.

Integrate systems of systems using AI at the edge. The next level of challenge for AI at the edge involves near-real-time interactions of multiple large distributed systems. For example, industry is working on autonomous vehicles, while the research community is thinking about a future with smart vehicles fully integrated with smart transportation infrastructure and smart cities [14]. Additional mobility players are on the horizon, from autonomous aerial and ground devices to interactions with wearables associated with pedestrians and cyclists. Indeed, AI at the edge will be ubiquitous in cities and mobility systems, making it extremely difficult to centralize the necessary information from tens of thousands of independent AI-controlled devices to understand their emergent behaviors. A distributed AI-at-the-edge approach is the only feasible solution to address this need to integrate multiple distributed systems. Each of the participating systems needs to be open and interoperable, and sophisticated edge AI capabilities will perform tasks such as negotiation and optimization across many interacting AI devices and services, as well as detection and prevention of failures due to system interactions, natural events, or intentional attacks—all the while balancing necessary data exchange with personal privacy. Such distributed AI capability would also be critical for improving the reliability of the nation's electric power grid, oil production, and other energy-related systems.

3. Advances in the Next Decade

As AI further permeates everyday life over the next decade, industry will turn to low-power edge devices for AI computation. The current paradigm of sending data back to a data center for analysis will no longer be tractable as the volume of data being collected becomes too large and the speed upon which it needs to acted increases with the need for real-time control. Industry has invested heavily in a variety of edge computing devices for AI, including tensor calculation accelerators (e.g., Google's Edge TPU and Intel's Movidius) and neuromorphic devices (e.g., IBM's TrueNorth and Intel's Loihi). There will be dramatic improvements in the power consumption and compute capability of these devices over the next decade.

Industries are also at the forefront of developing streaming data analysis systems and data standards. For example, many companies such as Waymo [15], Tesla [16], and Uber [17] are developing various versions of the self-driving software platform to go with their own vehicles. Many of the general-purpose ML systems are also creating streaming versions for mobile and embedded applications (e.g., Google TensorFlow has TensorFlow Lite, and PyTorch has a number of distributed backend systems that could support embedded applications).

It is anticipated that the commercial technologies will progress quickly in the next decade; however, these advancements are unlikely to meet the needs of the grand challenges mentioned in the previous sections. Various scientific applications will create data much faster than commercial applications. For example, monitoring the environment and the electric power grid will require data integration on a much larger scale than any commercial

enterprise. The connected mobility systems may have data rate, data volume, and data variety challenges not seen in commercial uses. Even for hydraulic fracturing applications, where there is clear commercial interest, there might still be the need for DOE or some other agencies to fund the initial development of the data analysis technology as in the case of the drill head used for hydraulic fracturing [18,19]. Funding the underlying data analysis research will benefit many applications, directly or indirectly, with far-reaching impacts.

4. Accelerating Development

Supporting AI at the edge will be very important to many future DOE efforts. Much as AI is permeating everyday life, it is also permeating nearly every field of science. This often takes place as an analysis of static data sets collected in advance. However, the ability to analyze data as it is collected or to exert real-time control over experiments presents an incredible opportunity to achieve discoveries that otherwise would not be possible.

The grand challenges mentioned previously may be able to leverage industrial development to a certain extent. For example, the challenge of handling data from high-speed sensors may be resolved by leveraging the new AI hardware and more computing capability per watt. However, the need to integrate systems of systems is unlikely to be addressed by industry. Therefore, to accelerate development, the following key algorithmic and mathematical challenges, derived from distinct application requirements, will have to be addressed. (Note that some of these topics overlap with those described in Chapters 7–9.)

Learning under limited resources. Edge computing is expected to operate under a number of resource limitations. For example, the computing resources at the edge are likely to be much smaller than could be available at a cloud data center or HPC center. Because of this, AI at the edge is going to work with limited data, presumably only the most recent data records. Under such limitations, the AI model is likely to be relatively small in size and will have to be updated periodically to accommodate new trends. New algorithms will be needed to cope with such resource constraints.

Understanding errors, failures, and correctness. Devices at the edge are often unreliable, and the data collected could be noisy or otherwise imperfect. Correctly understanding the impact of the noise, errors, and failures on data analysis operations and control actions will be another challenging issue. More broadly, improving reliability, robustness, and interpretability is a key fundamental research topic for AI.

Dealing with all aspects of the computing continuum. To bring the promise of AI at the edge to DOE science domains, there is a need to smoothly connect the edge and the core. Currently, there is no unified programming framework for the “computing continuum”—storage, networking, and computing resources from edge to fog to cloud. A better way to describe, model, and program the computing continuum from components and behaviors to systems, objectives, and intents is needed.

Modeling interactions. For edge devices to properly interact with the core and other edge devices, information and AI models have to be exchanged and understood by all parties involved. Modeling the interactions is critical to allow larger systems to be composed from individual components and smaller systems. Limited work is currently available on this topic.

Managing dynamic resources and interacting systems. AI at the edge requires new edge-focused resource management and support for multitenancy. Current edge computing systems, such as Waggle, concentrate on a single device; future edge nodes must be able to support multiple AI workloads scheduled to match sampling rates or operational needs. Additionally, the diverse nature of edge computing requires heterogeneity of edge computing hardware.

Research on how to design and optimize these heterogeneous computing nodes in a systematic and scientific manner is needed.

5. Expected Outcomes

DOE's mission demands high-performance AI at the edge to harness the power of large experiments and supercomputers. The anticipated work will allow data collection and analyses at scales not possible in the absence of edge computing. By investing in highly capable, robust, and versatile edge devices, DOE will enable scientists to perform large-scale experiments in harsh environments. AI at the edge will empower scientists to modify their experiments in real time based on the data being collected and thus drive them toward discoveries that would not be achievable otherwise. High-performance AI at the edge will fundamentally change the way DOE scientists work.

6. References

1. "Edge Computing: Vision and Challenges," June 9, 2016 (https://ieeexplore.ieee.org/document/7488250, accessed October 11, 2019).
2. "Edge-centric Computing–DOIs," September 30, 2015, http://doi.org/10.1145/2831347.2831354, accessed October 11, 2019.
3. "Patterned Probes for High Precision 4D-STEM Bragg Measurements," July 11, 2019, https://arxiv.org/abs/1907.05504, accessed October 11, 2019.
4. "Making the Invisible Visible: New Sensor Network Reveals Telltale Patterns in Neighborhood Air Quality," July 22, 2019, https://newscenter.lbl.gov/2019/07/22/new-sensor-network-neighborhood-air-quality/, accessed October 11, 2019.
5. "Waggle: An open sensor platform for edge computing," https://ieeexplore.ieee.org/abstract/document/7808975/, accessed October 11, 2019.
6. "Array of things: a scientific research instrument in the public way," April 18, 2017, https://dl.acm.org/citation.cfm?id=3063771, accessed October 11, 2019.
7. "Argonne supports grid advances through pioneering energy storage and sensor research," March 18, 2019, https://www.anl.gov/es/article/argonne-supports-grid-advances-through-pioneering-energy-storage-and-sensor-research, accessed October 11, 2019.
8. "Edge TPU–Google Cloud," https://cloud.google.com/edge-tpu/, accessed October 11, 2019.
9. "Intel Movidius, an Intel Company," https://www.movidius.com/, accessed October 11, 2019.
10. "Brain-inspired Chip–IBM Research," http://www.research.ibm.com/articles/brain-chip.shtml, accessed October 11, 2019.
11. "AR1K–The Smart Farm Research Consortium," https://ar1k.org/, accessed October 11, 2019.
12. "AR1K: Sustainable, Profitable Agriculture through Research," https://eesa.lbl.gov/projects/ar1k-sustainable-profitable-agriculture-research/, accessed October 11, 2019.
13. "ESnet," http://es.net/, accessed October 14, 2019.
14. "Smart Cities: The Future of Urban Development," *Forbes*, May 19, 2019, https://www.forbes.com/sites/jamesellsmoor/2019/05/19/smart-cities-the-future-of-urban-development/, accessed October 14, 2019.
15. "Waymo," https://waymo.com/, accessed October 14, 2019.
16. "Tesla," https://www.tesla.com/, accessed October 14, 2019.
17. "Earn Money by Driving or Get a Ride Now," https://www.uber.com/, accessed October 14, 2019.

18. “Hydraulic fracturing Sandia’s role in shale gas production technologies,” http://energy.sandia.gov/wp-content/gallery/uploads/FINAL-HydraulicFracturing-Final-wSAND1.pdf, accessed October 14, 2019.
19. “Hydraulic Fracturing: A Public-Private R&D Success Story,” https://clearpath.org/energy-101/hydraulic-fracturing-a-public-private-rd-success-story/, accessed October 14, 2019.

16. Facilities Integration and AI Ecosystem

Recent advances in AI have been driven by the ability to collect, store, and process large labeled datasets using large HPC and HPN facilities. DOE HPC facilities represent some of the world's largest computational and data ecosystems for generating, moving, and analyzing experimental and simulation data. These facilities are uniquely positioned to be centers for advances in AI research and applications and must therefore be prepared to fully support these capabilities in the next decade. Improving integration among DOE user facilities will ensure scientists have what is needed to apply AI methods in their research.

1. State of the Art

Data Management and Movement: Access to Data. AI derives its effectiveness from statistical generalizations gleaned from large volumes of often high-dimensional data. Within the scientific community, such data can be found at experimental, observational, and computational facilities, and the path forward requires making it readily available for use with AI applications. Most data management challenges have been described by what are known as the FAIR data principles. However, the infrastructure for managing, curating, publishing, and cataloging datasets that adhere to these principles has yet to reach the same level of maturity as the storage, compute, and network infrastructures.

Resource Orchestration: Co-scheduling and Co-designing. Practically all scientific domains are undergoing paradigm shifts due to explosions in the volume, variety, and velocity of datasets. Effective exploration of data via AI methods necessitates the tight coupling of experimental, observational, computational, and data facilities within and beyond the DOE complex. The tight coupling includes seamless access (i.e., authorization, authentication) and consistent interfaces to facilities regardless of location, HPNs to connect facilities, and co-scheduling of resources. Much of the work to date has been at the prototype level and additional functionality, including resource modeling, resource scheduling, and trust models, is still needed.

Smart Facilities: AI to Enable AI. HPC and HPN are capable of generating a comprehensive range of operational statistics with potential to leverage AI capabilities for facility control, monitoring, and management. For example, the scientific community is exploring the use of AI models on operational and application data generated by facilities to identify and proactively predict hardware failures before they occur. Standards for data collection, data and metadata representation, and data curation have not yet been established, presenting opportunities to exploit AI capabilities and increase the operational efficiency of the large HPC ecosystems deployed by DOE.

2. Major (Grand) Challenges

The overarching challenge for realizing the full potential of data-driven science is the development of the infrastructure required to facilitate AI applications. At least three major challenges have emerged in the quest for comprehensive facilitation of AI workloads.

Enable greater access to data. For years, scientists have decried the rate of growth of scientific information (estimated in one recent study to double roughly every 9 years*). Thus, data management will present a major challenge to the application of AI for science research (see Chapter 12, Data Life Cycle and

* Richard Van Noorden, "Global scientific output doubles every nine years," *Nature News Blog*, May 7, 2014, (http://blogs.nature.com/news/2014/05/global-scientific-output-doubles-every-nine-years.html).

Infrastructure). This includes developing a broad range of capabilities, such as software and services for accessing, sharing, managing, searching, discovering, publishing, cataloging, and curating data, in addition to tracking provenance and community-driven best practices for representing, storing, and exchanging data.

Facilities will need to collaborate to develop a unified scientific data management system that simplifies routine operations like searching, organizing, sharing, moving, and annotating data via an uncomplicated user interface. Simplified access to data from a heterogeneous collection of facilities, through technologies such as federated identities, will be key. Such a system will need to account for—and, if necessary, abstract—different security and authentication protocols used at various facilities and be able to offer necessary security for handling sensitive data associated with national security or health applications. Users should not be concerned with which file system or repository will be used to host their newly published data. The system will need to interface with software environments, computational workflows, and scientific instruments to extract relevant parameters, calibration information, and source datasets relevant to downstream consumers of the data products (see Chapter 11, Software Environments and Software Research).

A central data management system needs to be closely linked with ancillary services that would handle specific operations such as publishing, cataloging, and curating datasets. While many such stand-alone capabilities exist today, these systems do not communicate well with each other to form an integrated, strongly knit family of services. These data systems need to be designed with long-term storage and routine movement of petabyte-sized datasets in mind. The ability to index millions to billions of data records across a facility or combination of facilities will also be required. The division of larger scientific data repositories into smaller data silos may be inevitable due to differences in how they are generated including programmatic expectations. These silos, however, will need to interoperate with the central data management service in a way that provides a consistent interface to users.

In addition to developing the necessary data infrastructure, facilities, publishers, and sponsors will need to collaboratively develop policies to encourage best practices. These include publishing and referencing datasets used for journal publications, the use of community-driven and open source file formats to exchange data, and the use of community-standard schema and vocabulary to express rich metadata that describes and provides context to published data. Scientific domains will need to agree on data standards that include documentation on community access to facility data archives. Facilities may be able to play a role in organizing scientific domain–oriented workshops where experts develop domain-specific metadata and data standards that are compliant with the underlying data services.

In 10 to15 years, an ecosystem of connected facilities and networks will be needed to host, curate, and share domain training datasets and current state-of-the-art trained models for the scientific community. At the core of this ecosystem will be facility-specific, metadata-rich data catalogs with programmatic interfaces that enable the automation of data discovery, data movement, and AI training.

Develop AI-focused HPC hardware. Facilitating AI application science will be a substantial change for all facilities as they will need to broaden support beyond traditional modeling and simulation work to include observational and experimental components. This will include complex data interactions that lead to new scientific opportunities. Specifically, future HPC facility architectures will need to be better optimized to handle more complex data traffic, both within the facility and with external facilities of all types (see

Chapter 13, Hardware Architectures). HPC centers will need to pay special attention to high-performance networking for routinely moving the petabyte-scale datasets required for AI training tasks that will include file system I/O performance. The architectural details touch all parts of the computing ecosystem and will need to be systematically optimized to deliver the highest scientific impact.

In contrast to traditional simulation-only workloads whose data footprints typically have small inputs but large structured outputs, AI workloads will need to access large volumes of unstructured data, sometimes repeatedly, with a need to also write relatively modest checkpoint and output files. File systems can better support these workloads with faster random read speeds while compute nodes would benefit from fast burst buffers to maintain a local cache of frequently used input data. High-speed interconnects will need to facilitate fast periodic ALL-to-ALL random exchange of training data and gradient synchronization. AI workloads may also have large memory footprints. Because AI models can be trained with single and even half-precision arithmetic (unlike simulations, which often need double precision), CPU architectures capable of supporting operations with varying precisions will help maximize throughput in AI workloads. Similarly, future high-performance computers could use next-generation accelerators such as TPUs or neuromorphic computing units that are better suited for AI workloads as extensions for more conventional CPU and GPU architectures.

On a broader scale, data may need to be transferred across the ESnet HPN and may need to leverage computing and specialized AI hardware close to the instrument or sensor distributed across the network. This edge computing need could apply to data streamed directly through multiple stages of a network where it is used and then discarded. Every link in this chain—data portals, networks, edge computation, HPCs, and I/O systems—needs to be architected with AI applications in mind to efficiently exploit the huge potential gains in distributed computational performance.

Facilitate resource orchestration. Ensuring that data can be brought to a heterogeneous, distributed compute infrastructure needed for AI-based science requires new levels of cross-facility coordination and orchestration. AI workflows can include a variety of components, such as experimental data, multiple data repositories, local and nonlocal computing platforms, LANs and WANs, storage, and people in the loop. Policies at facilities will need to be restructured to allow seamless co-scheduling of these heterogeneous resources for scientific productivity. The federation of disparate or geographically distant facilities will need to be overcome through the development of standardized protocols and cross-facility identity management to allow movement of computational workloads and data. For example, these advancements will be critical for enabling “self-driving experimental facilities” (see Chapter 14, AI for Imaging).

Additionally, the orchestration of resources varies at different time scales and may be neither aligned nor readily predictable. Experiments may need quasi-real-time, AI-powered data analyses that depend on experimental operating and downtime schedules. In contrast, multiple runs of an AI training algorithm to find optimal network hyperparameters may tolerate a multiday turnaround period. Resource orchestration across multiple facilities will need to account for the urgency of the request, which may require high-priority, on-demand computing for some cases.

A global, AI-driven resource orchestrator will need to account for the heterogeneous computing landscape, as each computing site will have unique capabilities. The AI resource orchestrator could direct data and computational resources based on the optimal path and location for hardware, availability, energy costs, and the specific scientific application. Not only would this result in a

more efficient workload for the scientist, but there is potential to make more efficient use of network and computational facilities, avoiding bottlenecks and maintaining high use of all resources. Such an orchestrator could provide continuous feedback on how to improve efficiency and performance.

Leverage AI to enable AI with smart facilities. The increasing complexity of HPC and HPN workloads will require innovations in facility operations, and AI will play a critical role in driving this evolution. AI workloads present unique challenges because of their data movement patterns and uncommon mix of compute intensity and I/O (e.g., training vs. inference). In the short term, it will be important to develop representative AI benchmarks to characterize AI workflows and understand the optimizations required to efficiently support workflows associated with these use cases. This will involve developing AI benchmarks that expose operational data from the facilities through exemplar training and inference workloads. This information can then be used to build the tools and infrastructure to support AI at scale across the DOE complex, where these AI benchmarks should become an integral part of those currently used by computational and networking facilities.

A long-term goal for facility operations would be to drive operational decision-making using AI methods. A truly automated, optimized facility will be able to predict faults, detect anomalies or performance degradation, and balance the computational workload accordingly. However, for this grand challenge to be met, the right operational data need to be identified and collected. Identifying the dataset of telemetry that can be used by researchers to design autonomous behavior is not a trivial task—facilities currently produce numerous terabytes of telemetry data per day on everything from network statistics to power consumption. Identifying, curating, cleaning, and sharing these data are vital to designing a truly automated facility, as well as to developing a smart resource orchestrator. AI could also be used to simplify or automate access to software modules, input datasets, validate computational or experimental runs against previous runs, recommend new parameters for runs to avoid duplication, and study unexplored phenomena.

Meeting this challenge will enable AI to tune the ecosystem to create a more effective environment for AI applications. This will be of general benefit to users of such facilities due to the more flexible and performance-based environment. In addition, this will be of huge value to the facilities themselves, empowering them to predict usage patterns, identify trends in resource use, and make more informed decisions about future architectures (see Chapter 9, AI for Computer Science).

3. Advances in the Next Decade

The next three years will see the deployment of ESnet6 and NERSC-9 (Perlmutter), and the first generation of exascale machines (Aurora at ALCF and Frontier at OLCF) across the DOE complex (Figure 16.1).

All of these facilities already support the most popular AI frameworks, and it is expected that DOE will support the development of additional HPC-focused AI frameworks in the next decade, along with platforms that facilitate sharing and publishing AI networks, hyperparameters, and weights in a framework-agnostic and architecture-agnostic manner. The DOE ASCR facilities are all developing programs to increase support of AI. These programs will foster burgeoning AI applications in HPC ecosystems and need to be folded into generalized allocation programs. The deployment of scalable scientific data management systems that will form the foundation for curating high-quality datasets will also be necessary. This work will continue with the deployment of data gateways that facilitate the transfer of data from a variety of sources to computational facilities. It is also expected that AI will be extended to support rapid data processing at HPC

Figure 16.1 *Over the next three years, DOE will stand up its first generation of exascale machines. These systems, along with the upcoming ESnet high-performance network, present a unique opportunity to leverage HPC in the development of AI for science.*

facilities to enable quasi-real-time feedback on experiments and observations. The data gateway and the scientific data management system will be critical components expected to substantially reduce the accumulation of "dark" (i.e., unpublished) data and accelerate the accumulation of well-annotated and standardized data for AI in the upcoming decade.

Looking further ahead, the ASCR facilities will continue to design complex, technically advanced networking and computing facilities for future science generations where the needs of the AI ecosystem will be an integral part of any initial design. Given the pace of change in AI technology and techniques, these future facilities will also need to be designed with flexibility in mind to take advantage of the advances that will inevitably come from application work over the next decade.

4. Accelerating Development

An AI agent is only as capable as the quality of data used to train it. Currently, we lack the infrastructure and policies to facilitate curation of the high-quality datasets critical to fully realizing the potential of AI. The FAIR data principles provide ample guidelines for reaching this goal. Facilitating AI necessitates additional manpower for the further development of data management, movement, curation, publication, standardization, and streaming software/services (see Chapter 12, Data Life Cycle and Infrastructure). Some work has already begun along these lines at every DOE facility. However, a highly coordinated effort across the DOE complex will be necessary for rapid progress in this area. Software and services can facilitate good data practices that will feed AI agents, but actual accumulation of high-quality datasets is contingent on researchers using the aforementioned data software stack to populate data repositories. Policies must be developed to minimize generation of dark data and maximize generation of well-annotated data. AI efforts will be necessary to draw insights from the collected data, but facilities need to first train their researchers on ML, including DL, techniques. Furthermore, facilities will need to foster AI development through dedicated research programs. Given the data explosion in practically all scientific domains, facilities will need to train researchers on using high-performance computers for developing, scaling, and deploying AI agents that can leverage the ballooning body of data.

5. Expected Outcomes

Without the support of DOE facilities, the scientific community will struggle to take advantage of the promise of AI. The processing power of DOE supercomputers, including the forthcoming exascale systems, is vital to train AI algorithms using the huge amounts of data being produced and curated by the scientific community. However, simply building these computing facilities does not guarantee that they will be accessible and useful for AI research. The infrastructure described in this chapter will be essential to allow scientists to take full advantage of the compute resources DOE offers. AI will itself be essential to creating such an infrastructure.

With appropriate direction, funding, and the cross-facility cooperation described in this chapter, the goal of a seamlessly interconnected DOE complex can be achieved in 10 years. Such a reality will allow scientists to build AI-driven experimentation and discovery workflows, optimized and controlled by embedded AI in a transparent facility infrastructure spanning the DOE complex, allowing data and compute resources to be directed according to the needs of the scientists and the availability of resources, without a human in the loop.

AA. Report Writing Team

The following individuals made significant contributions to the final content of this report.

First Name	Last Name	Institution
Corey	Adams	Argonne National Laboratory
Srikanth	Allu	Oak Ridge National Laboratory
Jim	Ang	Pacific Northwest National Laboratory
Mihai	Anitescu	Argonne National Laboratory
Katerina	Antypas	Lawrence Berkeley National Laboratory
Melina L.	Avila Coronado	Argonne National Laboratory
Prasanna	Balaprakash	Argonne National Laboratory
Deborah	Bard	Lawrence Berkeley National Laboratory
Pete	Beckman	Argonne National Laboratory
Wes	Bethel	Lawrence Berkeley National Laboratory
Philip	Bingham	Oak Ridge National Laboratory
Kristofer	Bouchard	Lawrence Berkeley National Laboratory
Thomas S.	Brettin	Argonne National Laboratory
Ben	Brown	Lawrence Berkeley National Laboratory
Paolo	Calafiura	Lawrence Berkeley National Laboratory
Charles E.	Catlett	Argonne National Laboratory
Andrew	Chien	Argonne National Laboratory
Taylor	Childers	Argonne National Laboratory
Santanu	Chaudhuri	Argonne National Laboratory
Ian C.	Cloet	Argonne National Laboratory
Ren	Cooper	Lawrence Berkeley National Laboratory
David	Dean	Oak Ridge National Laboratory
Bert	deJong	Lawrence Berkeley National Laboratory
Marcel	Demarteau	Oak Ridge National Laboratory
Sudip	Dosanjh	Lawrence Berkeley National Laboratory
Dipankar	Dwivedi	Lawrence Berkeley National Laboratory
Nicola	Ferrier	Argonne National Laboratory
Ian	Foster	Argonne National Laboratory
Hector	Garcia Martin	Lawrence Berkeley National Laboratory
Devarshi	Ghoshal	Lawrence Berkeley National Laboratory
Chin	Guok	Lawrence Berkeley National Laboratory
Dogan	Gursoy	Argonne National Laboratory
Salman	Habib	Argonne National Laboratory
James	Hack	Oak Ridge National Laboratory
Kawtar	Hafidi	Argonne National Laboratory
Kenneth	Herwig	Oak Ridge National Laboratory
Judith	Hill	Oak Ridge National Laboratory
Forrest M.	Hoffman	Oak Ridge National Laboratory
Tianzhen	Hong	Lawrence Berkeley National Laboratory
David	Humphreys	General Atomics
Barbara	Jacak	Lawrence Berkeley National Laboratory

First Name	Last Name	Institution
Cynthia	Jenks	Argonne National Laboratory
Mariam	Kiran	Lawrence Berkeley National Laboratory
Rao	Kotamarthi	Argonne National Laboratory
Ana	Kupresanin	Lawrence Livermore National Laboratory
Teja	Kuruganti	Oak Ridge National Laboratory
Frank	Liu	Oak Ridge National Laboratory
Bronson	Messer	Oak Ridge National Laboratory
Zein-Eddine	Meziani	Argonne National Laboratory
Georgios	Michelogiannakis	Lawrence Berkeley National Laboratory
Inder	Monga	Lawrence Berkeley National Laboratory
Dmitriy	Morozov	Lawrence Berkeley National Laboratory
Peter	Nugent	Lawrence Berkeley National Laboratory
Michael E.	Papka	Argonne National Laboratory
Mary Ann	Piette	Lawrence Berkeley National Laboratory
Alan	Poon	Lawrence Berkeley National Laboratory
Prabhat		Lawrence Berkeley National Laboratory
Brian	Quiter	Lawrence Berkeley National Laboratory
Lavanya	Ramakrishnan	Lawrence Berkeley National Laboratory
Nageswara	Rao	Oak Ridge National Laboratory
Rob	Ross	Argonne National Laboratory
Rajesh	Sankaran	Argonne National Laboratory
Jibo	Sanyal	Oak Ridge National Laboratory
Martin	Schoenball	Lawrence Berkeley National Laboratory
Koushik	Sen	University of California, Berkeley
John	Shalf	Lawrence Berkeley National Laboratory
Arjun	Shankar	Oak Ridge National Laboratory
Michael	Smith	Oak Ridge National Laboratory
Suhas	Somnath	Oak Ridge National Laboratory
Bobby G.	Sumpter	Oak Ridge National Laboratory
Georgia	Tourassi	Oak Ridge National Laboratory
John	Turner	Oak Ridge National Laboratory
Tom	Uram	Argonne National Laboratory
James	Vary	Iowa State University
Velimir (Monty) V.	Vesselinov	Los Alamos National Laboratory
Jeffrey	Vetter	Oak Ridge National Laboratory
Venkatram	Vishwanath	Argonne National Laboratory
Haruko	Wainwright	Lawrence Berkeley National Laboratory
Stefan	Wild	Argonne National Laboratory
David	Womble	Oak Ridge National Laboratory
John	Wu	Lawrence Berkeley National Laboratory
Junqi	Yin	Oak Ridge National Laboratory
Steven	Young	Oak Ridge National Laboratory
Piotr	Zarzycki	Lawrence Berkeley National Laboratory
Petrus	Zwart	Lawrence Berkeley National Laboratory

AB. Agendas

AI for Science Town Hall
Argonne National Laboratory
Advanced Photon Source (APS), Building 402
July 22–23, 2019

Monday, July 22, 2019

8:30 a.m. **Registration……..APS Main Lobby**

9:00 a.m. **Welcome……APS Auditorium**
Kim Sawyer

9:10 a.m. **Introductory Remarks……APS Auditorium**
Congressman Bill Foster

9:20 a.m. **Opening Statement……..APS Auditorium**
Barbara Helland

9:30 a.m. **AI for Science Opportunities……APS Auditorium**
Rick Stevens

10:30 a.m. **Morning Break**

10:45 a.m. **AI at Scale 1: Cosmology……..APS Auditorium**
Salman Habib

11:05 a.m. **AI at Scale 2: Materials……..APS Auditorium**
Ian Foster

11:25 a.m. **AI at Scale 3: Climate……APS Auditorium**
Rao Kotamarthi

11:45 a.m. **Breakout Session Charge Questions……...APS Auditorium**
Rick Stevens

12:00 p.m. **Collect Lunch and Proceed to Application Breakout Sessions**

Materials, Chemistry and Nanoscience……..TCS 1404/1405
Co-leads: Cynthia Jenks, Tim Germann
Session scribe: Chris Knight

Materials, Chemistry and Nanoscience……..…TCS 1406/1407
Co-leads: Steve Plimpton, Pieter Swart
Session scribe: Huihuo Zheng

Imaging and Scientific User Facilities……APS Gallery
Co-leads: Nicola Ferrier, Shinjae Yoo
Session scribe: Nicholas Schwarz

Imaging and Scientific User Facilities……..….APS E1100/E1200
Co-leads: Barry Chen, Christine Sweeney
Session scribe: Justin Wozniak

Environment, Climate and Earth Science…….APS A1100
Co-leads: Rao Kotamarthi, Haruko Wainwright
Session scribe: Scott Collis

Biology and Life Science…….APS Auditorium
Co-leads: Thomas S. Brettin, Ben Brown
Session scribe: Gyorgy Babnigg

Fundamental Physics……..….TCS 1416a
Co-leads: Katrin Heitmann, Paolo Calafiura
Session scribe: Corey Adams

Engineering and Technology……..Bldg. 241 D172
Co-leads: Santanu Chaudhuri, Stuart Slattery
Session scribe: Shashikant Aithal

Energy (wind, solar, fossil, etc.)…….TCS 1416b
Co-leads: Mihai Anitescu, Bill Tang
Session scribe: Julie Bessac

2:40 p.m. **Breakout Sessions End**

3:00 p.m. **Breakouts Report Out (10 minutes each)……..….APS Auditorium**

4:30 p.m. **Day One Close-out Summary……...APS Auditorium**
Rick Stevens

5:00 p.m. **Adjourn**

Tuesday, July 23, 2019

8:30 a.m. **Registration……..APS Main Lobby**

9:00 a.m. **Summary of Day 1 and Day 2 Cross-cut Charge……....APS Auditorium**
Rick Stevens

9:30 a.m. **Technological and Cross-cut Breakout Sessions**

Optimization / UQ / Statistics……..……..TCS 1404/1405
Co-leads: Stefan Wild, Clayton Webster
Session scribe: Bethany Lusch

Optimization / UQ / Statistics……..……..TCS 1406/1407
Co-leads: Ana Kupresanin, Earl Lawrence
Session scribe: Vishwas Rao

Convergence of Simulation and Data Methods……...TCS 1416a
Co-leads: Emil Constantinescu, Frank Alexander
Session scribe: Taylor Childers

Convergence of Simulation and Data Methods……...TCS 1416b
Co-leads: Justin Newcomer, Cory Hauck
Session scribe: Hong Zhang

Data Infrastructure and Life Cycle……PS E1100/E1200
Co-leads: Ian Foster, Kerstin Kleese van Dam
Session scribe: Youssef Nashed

Hardware and Architecture……APS Gallery
Co-leads: Andrew Chien, Jeffrey Vetter
Session scribe: Murali Emani

Software Environments and Software Research…….APS Auditorium
Co-leads: Prasanna Balaprakash, Devarshi Ghoshal
Session scribe: Tom Uram

Facilities Integration…….APS A1100
Co-leads: Michael E. Papka, Arjun Shankar
Session scribe: Ryan Milner

11:40 a.m. **Collect Lunch and Proceed to Report Out Session**

12:00 p.m. **Breakouts Report Out (10 minutes each)……..APS Auditorium**

1:30 p.m. **Town Hall Close-out with Next Steps…….APS Auditorium**
Rick Stevens

3:00 p.m. **Town Hall Concludes**

AI for Science Town Hall
Oak Ridge National Laboratory
ORNL Conference Center
August 20–21, 2019

Tuesday, August 20, 2019

8:00 a.m. **Registration and Working Continental Breakfast……..….ORNL Conference Center**

8:30 a.m. **Welcome and Introduction……..…ORNL Conference Center**
Jeffrey Nichols

8:35 a.m. **ORNL Opening Remarks ……..……..ORNL Conference Center**
Jeff Smith

8:45 a.m. **DOE HQ Opening Remarks…….……..…ORNL Conference Center**
Steve Binkley

9:00 a.m. **Keynote: AI for Science Opportunities…….....ORNL Conference Center**
David Womble

9:40 a.m. **Plenary Session……ORNL Conference Center**
Session Chair: *Doug Kothe*

AI at Scale 1: Microscopy
Sergei Kalinin

AI at Scale 2: Advanced Manufacturing
Tom Kurfess

AI at Scale 3: Health
Georgia Tourassi

10:40 a.m. **Breakout Session Charge Questions……...ORNL Conference Center**
Jeffrey Nichols

11:00 a.m. **Collect Lunch and Proceed to Application Breakout Sessions**

Materials, Chemistry and Nanoscience……..…Tennessee B
Co-Leads: Bobby G. Sumpter, Markus Eisenbach, Wibe de Jong

Data Collection, Reduction, Analysis, and Imaging for Scientific User Facilities…….……..…...Tennessee C
Co-Leads: Hans Christian, Sean Hearne, Christine Sweeny, Jack Wells, Thomas Proffen

Environment, Climate and Earth Science……..…Tennessee A
Co-Leads: Forrest M. Hoffman, Alison Boyer, Velimir (Monty) V. Vesselinov

Biology and Life Science…….……Emory
Co-Leads: Julie Mitchell, Jacob Hinkle, Ben Brown

Fundamental Physics…….……Cumberland
Co-Leads: Marcel Demarteau, Bronson Messer, Torre Wenaus

Fusion Energy……..…….Building 5700, Room F234
Co-Leads: Phil Ferguson, Mike Churchill, John Canik

Transportation and Mobility……..5700, CASL Room B302a
Co-Leads: Robert Wagner, Jibo Sanyal, Stanley Young

Advanced Manufacturing……..Building 5600, EVEREST (B228)
Co-Leads: Stuart Slattery, Vincent Paquit, Jim Belak

Energy Generation & Distribution……Building 5700, Room L204
Co-Leads: Teja Kuruganti, Tara Pandya, Mike Sprague

3:00 p.m. **Breakout Reports Out (10 minutes each) .....ORNL Conference Center**

4:30 p.m. **Day One Close-out Summary.....ORNL Conference Center**
Jeffrey Nichols

5:00 p.m. **Reception .....ORNL Conference Center Lobby**

Wednesday, August 21, 2019

8:00 a.m. **Registration and Working Continental Breakfast……..ORNL Conference Center**

8:30 a.m. **Day 2 Welcome ……..ORNL Conference Center**
Thomas Zacharia

8:45 a.m. **Summary of Day 1 and Day 2 Cross-cut Charge……ORNL Conference Center**
Jeffrey Nichols

9:00 a.m. **Technological and Cross-cut Breakout Sessions**

Numerical Aspects of Learning…….……...Building 5700, Room F234
Co-Leads: Clayton Webster, Stefan Wild, Sandeep Madireddy

Model Applicability and Characterization……Tennessee B
Co-Leads: Blair Christian, Dan Lu, Justin Newcomer

Decision Support ……..5700, CASL Room B302a
Co-Leads: Rick Archibald, Tom Potok, Cynthia Phillip

Science Informed Learning ……..……...…Tennessee C
Co-Leads: Scott Klasky, Cory Hauck, Jeff Hittinger

Software Environments and Software Research…..……...Emory
Co-Leads: Robert Patton, Judith Hill, Eric Cyr

Data Infrastructure & Life Cycle……..Tennessee A
Co-Leads: Arjun Shankar, Katie Knight, Brad Settlemyer

Hardware and Architecture……..……..Building 5600, EVEREST (B228)
Co-Leads: Katie Schuman, Travis Humble, Kenneth Alvin

Facilities Integration and AI Ecosystem …….....Cumberland
Co-Leads: James Hack, Michael E. Papka, Inder Monga

12:00 p.m. **Collect Lunch and Head Back to Breakout Session**

1:00 p.m. **Final Report Out from Breakout Session (10 minutes each)**

2:30 p.m. **Town Hall Close-out with Next Steps……..…ORNL Conference Center**
Jeffrey Nichols

3:00 p.m. **Town Hall Concludes**

AI for Science Town Hall
Lawrence Berkeley National Laboratory
Building 50 Auditorium
September 11–12, 2019

Wednesday, September 11, 2019

7:15 a.m. **Registration……Building 50 Auditorium Lobby**

7:30 a.m. **Networking Breakfast**

8:30 a.m. **Welcome and Introduction……..……..….Building 50 Auditorium**
Mike Witherell

8:40 a.m. **Opening Remarks……...Building 50 Auditorium**
Barbara Helland

8:50 a.m. **AI for Science Opportunities and Meeting Objectives**
Katherine Yelick

9:40 a.m. **Break**

9:55 a.m. **Examples of AI at Scale……....Building 50 Auditorium**
Session Chair: David Brown

AI, Machine Learning, and Experimental Facilities
James Sethian

AI at Scale: Astrophysics
Josh Bloom

AI at Scale in Biology
Ben Brown

11:25 a.m. **Breakout Logistics…….Building 50 Auditorium**
Katherine Yelick

11:30 a.m. **Collect Lunch and Proceed to Application Breakout Sessions**

11:45 a.m. **Application Breakout Sessions**

Physical Sciences
Coordinator: Paolo Calafiura

Cosmology and Astrophysics 59-4016
Co-leads: Uros Seljak, Salman Habib

Particle Physics 59-4022
Co-leads: Steve Farrell, Ariel Schwartzman

Accelerator Science 50-4058
Co-leads: Remi Lehe, Daniel Ratner

Fusion **59-4102**
Co-leads: CS Chang, Mike Zarnstorff

Energy Sciences
Coordinator: Jonathan Carter

Materials and Chemistry Modeling **66-316**
Co-leads: Anubhav Jain, Jeff Hammond, Shyam Dwaraknath

Materials Synthesis and Chemistry **62-203**
Co-leads: Carolin Sutter-Fella, Emory Chan, Ethan Crumlin

Light Sources **66-Aud**
Co-leads: Alex Hexemer, Petrus Zwart, Chris Jacobsen

Electron Microscopy Imaging **67-3111**
Co-leads: Mary Scott, Eva Nogales, Marcus Hanwell

Earth and Environmental Sciences
Coordinator: Trever Keenan, Dipankar Dwivedi

Climate and Carbon **84-318**
Co-leads: Trevor Keenan, Nathan Urban, Esmond Ng

Subsurface and Geoscience **74-104**
Co-leads: Martin Schoenball, Piotr Zarzycki, Andrew Stack

Water **74-324**
Co-leads: Dipankar Dwivedi, Hoshin Gupta, Grey Nearing

Biological and Life Sciences
Coordinator: Ben Brown

Microbiome and Environmental Biology **59-3049**
Co-leads: Paramvir Dehal, Jennifer Clarke

Synthetic Biology **59-3042**
Co-leads: Hector Garcia Martin, Peter St. John

Health **59-3025**
Co-leads: Kris Bouchard, Tina Hernandez-Boussard

Engineering and Infrastructure
Coordinator: Peter Nugent

Engineering and Manufacturing **70A-3377**
Co-leads: Stuart Slattery, Tarek Zohdi

Transportation / Mobility **59-3104**
Co-leads: Cy Chan, Timothy Berg

Urban **59-3104**
Co-leads: Mary Ann Piette, Peter Graf

SmartGrid **59-3104**
Co-leads: William Hart, Russell Bent

Computer Science
Coordinator: Katherine Yelick

AI Networking and Computing Facilities **59-3101**
AI for anomaly detection, cybersecurity, networking management, etc.
Co-leads: Mariam Kiran, Nageswara Rao, Lavanya Ramakrishnan

AI for Computer Hardware and Software **59-3101**
AI for architecture design, programming, etc.
Co-leads: Georgios Michelogiannakis, Koushik Sen

2:30 p.m. **Breakout Sessions End** ..... **Building 50 Auditorium**

3:00 p.m. **Lightning Breakouts Report Out (5 minutes each)** ..... **Building 50 Auditorium**

5:00 p.m. **Networking Reception** ..... **Building 59 Plaza**

6:00 p.m. **Adjourn**

Thursday, September 12, 2019

7:15 a.m. **Registration** ..... **Building 50 Auditorium Lobby**

7:30 a.m. **Networking Breakfast**

8:30 a.m. **Summary of Day 1 and Day 2 Cross-cut Charge** ..... **Building 50 Auditorium**
Katherine Yelick

8:45 a.m. **Travel to breakout locations**

9:00 a.m. **Technological and Cross-cut Breakout Sessions**

Math Foundations for AI
Coordinator: Tamara Kolda (SNL)

Performance Optimization of Deep Learning **59-3054**
Numerical and stochastic optimization, network design, hyperparameter search, network compression, parallelization
Co-leads: Aydin Buluc, Sherry Li

Foundations and Challenges of Deep Learning ***59-3049***
Numerical properties of DL, problems with generalization, understanding how it works and failure modes, theoretical considerations
Co-leads: Tamara Kolda, Tess Smidt

Opportunities and Foundations of Traditional ML **59-3025**
Regression, random forests, support vector machines, principal component analysis, clustering, optimization methods
Co-leads: Justin Newcomer, Ali Pinar

Reinforcement/Streaming learning for Decision Support / Control **59-3070**
Real-time control and decision-making, incorporating feedback
Co-leads: Mike Mahoney, Prabhat

ML for science problems with limited data **59-4101**
Bayesian methods, matrix completion, statistical sampling design
Co-leads: Jeremy Templeton, Janine Bennett

Science-informed learning **59-4102**
Physics/chemistry/biology-constrained, data integration
Co-leads: Juliane Mueller, Stefan Wild

Uncertainty Quantification **59-3104**
Co-leads: Habib Najm, David Barajas-Solano

Use of AI with Simulation ..... **50 Auditorium**
Co-leads: Marcus Day, Katherine Lewis

Software Environments and Research..... **54-Perseverance Hall**
How will we write AI software? Tensorflow, Pytorch, etc., and DOE-developed alternatives or improvements for science? What OS services, workflows, etc. are needed?
Co-leads: Dmitriy Morozov, Charles Tripp

Data Lifecycle **59-3101**
Data preparation, data sets, traditional analytics, de-noising, provenance, etc.
Co-leads: Wes Bethel, John Wu

Hardware Technology **50B-4205**
Centralized HPC, Edge Devices…
Co-leads: John Shalf, James Ang

Facilities Infrastructure and Integration; the AI Ecosystem **70A-3377**
I/O balance, on-demand computing, science gateways, networking
Co-leads: Inder Monga, Deborah Bard, Michael E. Papka

Cybersecurity and Privacy **59-4016**
Security of Cyber-physical systems, data privacy
Lead: Sean Peisert

11:30 a.m. **Collect Lunch and Proceed in to Report Out Session** .. **Building 50 Auditorium**

11:45 a.m. **Breakouts Report Out (5 minutes each)**..... **Building 50 Auditorium**

1:45 p.m. **Town Hall Close-out with Next Steps** ..... **Building 50 Auditorium**
Katherine Yelick

2:00 p.m. **Town Hall Concludes**

AI for Science Town Hall
Washington, DC
Renaissance DC - Downtown Hotel
October 22–23, 2019

Tuesday, October 22, 2019

7:30 a.m. **Registration and Working Continental Breakfast……...Ballroom Level Lobby**

8:30 a.m. **Welcome and Introduction……Grand Ballroom North**
Barbara Helland

8:45 a.m. **DOE HQ Opening Remarks……...Grand Ballroom North**
Chris Fall

9:00 a.m. **Summary from 3 Town Halls.…….……..Grand Ballroom North**
Katherine Yelick, Rick Stevens, Jeffrey Nichols

10:00 a.m. **Break**

10:30 a.m. **How Significant will AI be for the Energy Sector?.……Grand Ballroom North**
Quantifying progress and outlining signposts
Claire Curry, Bloomberg New Energy Finance

11:15 a.m. **AI Research Update: What's Going On Around ……Grand Ballroom North**
The World and Our Research Plans for Studying AI For Science
Earl Joseph, Hyperion Research

11:45 a.m. **Break for Working Lunch**
Networking and Preparation for Breakout Sessions

1:30 p.m. **Breakout Sessions**

Machine Learning Foundations and Open Problems……Grand Ballroom North
Co-Leads: David Womble, Stefan Wild, Prabhat

Facilities Integration and AI Ecosystem……...Meeting Room 3
Co-Leads: James Hack, Michael E. Papka, Sudip Dosanjh, Inder Monga

Earth and Environmental Sciences……..Meeting Room 6
Co-Leads: Forrest M. Hoffman, Rao Kotamarthi, Haruko Wainwright

Chemistry, Materials, and Nano Science…….Meeting Room 7
Co-Leads: Cynthia Jenks, Bert deJong

Engineering and Manufacturing…….Meeting Room 8
Co-Leads: John Turner, Santanu Chaudhuri, Peter Nugent

Nuclear Physics…….....Meeting Room 9
Co-Leads: David Dean, Zein-Eddine Meziani, Brian Quiter

Data Life Cycle and Infrastructure……Meeting Room 10
Co-Leads: Arjun Shankar, Nicola Ferrier, Wes Bethel

Support for AI for Experimental Facilities…….Meeting Room 16
Co-Leads: Kenneth Herwig, Dogan Gursoy, Petrus Zwart

3:15 p.m. **Break**

3:30 p.m. **Startup Innovations in AI Hardware……..….Grand Ballroom North**
Moderator: Rick Stevens
Andy Hock
Kunle Olukotun
Dale Southard

4:30 p.m. **Breakout Summary……..……..Grand Ballroom North**
Valerie Taylor, Arthur Barney Maccabe, David Brown

5:00 p.m. **Close-out for the Day…….……Grand Ballroom North**
Barbara Helland

Wednesday, October 23, 2019

7:30 a.m. **Registration and Working Continental Breakfast**

8:30 a.m. **Day 2 Welcome…….Grand Ballroom North**
Barbara Helland

8:45 a.m. **Breakouts**
AI for Computer Science……Grand Ballroom North
Co-Leads: Nageswara Rao, Prasanna Balaprakash. Lavanya Ramakrishnan

Biology and Life Sciences……..……..Meeting Room 3
Co-Leads: Georgia Tourassi, Thomas S. Brettin, Ben Brown

High Energy Physics……..…...Meeting Room 6
Co-Leads: Salman Habib, Paolo Calafiura

Smart Energy Infrastructure……Meeting Room 8
Co-Leads: Teja Kuruganti, Mihai Anitescu, Tianzhen Hong

Software Environments and Software Research…….Meeting Room 9
Co-Leads: Judith Hill, Rob Ross, Katerina Antypas

Support for AI at the Edge……..Meeting Room 10
Co-Leads: Steven Young, Pete Beckman, John Wu

Hardware Architectures…….....Meeting Room 16
Co-Leads: Jeffrey Vetter, Andrew Chien, John Shalf

10:15 a.m. **Break**

10:30 a.m. **DOE Headquarters Remarks……..……Grand Ballroom North**
Paul Dabbar

10:45 a.m. **Cross Agency AI Strategies……..……..Grand Ballroom North**
Moderator: Lynne Parker (OSTP)
DOE – Steve Binkley
DOD NSA Research – Adam Cardinal-Stakenas
NSF - Erwin Gianchandani
NIH - Susan Gregurick

11:45 a.m. **Breakout Summary……..Grand Ballroom North**
Valerie Taylor, Arthur Barney Maccabe, David Brown

12:15 p.m. **AI Killer Applications…….…..Grand Ballroom North**
Rick Stevens, Katherine Yelick, Jeffrey Nichols

1:00 p.m. **Wrap Up……..Grand Ballroom North**
Barbara Helland

1:15 p.m. **Working Lunch…….…….Ballroom Level Lobby**
Networking and Coordination of Town Hall Report

This page intentionally blank.

AC. Combined Town Hall Registrants

First Name	Last Name	Institution
Brook	Abegaz	Loyola University of Chicago
Gina	Adam	George Washington University
Corey	Adams	Argonne National Laboratory
Marc	Adams	NVIDIA Corporation
Ryan	Adamson	Oak Ridge National Laboratory
Adetokunbo	Adedoyin	Los Alamos National Laboratory
Vivek	Agarwal	Idaho National Laboratory
Greeshma	Agasthya	Oak Ridge National Laboratory
Jeffery	Aguiar	Idaho National Laboratory
Lars	Ahlfors	Microsoft Corporation
James	Ahrens	Los Alamos National Laboratory
Sachin	Ahuja	CNH Industrial
James	Aimone	Sandia National Laboratories
Shashi	Aithal	Argonne National Laboratory
Adeel	Akram	Uppsala University
Maksudul	Alam	Oak Ridge National Laboratory
Frank	Alexander	Brookhaven National Laboratory
Boian	Alexandrov	Los Alamos National Laboratory
Yuri	Alexeev	Argonne National Laboratory
Stephanie	Allport	Bloomberg
Srikanth	Allu	Oak Ridge National Laboratory
Jeff	Alstott	Intelligence Advanced Research Projects Activity
Ilkay	Altintas	University of California, San Diego
Kenneth	Alvin	Sandia National Laboratories
James	Amundson	Fermi National Accelerator Laboratory
Valentine	Anantharaj	Oak Ridge National Laboratory
James	Ang	Pacific Northwest National Laboratory
Mihai	Anitescu	Argonne National Laboratory
Dionysios	Antonopoulos	Argonne National Laboratory
Katerina	Antypas	Lawrence Berkeley National Laboratory
Chid	Apte	IBM Research
Rick	Archibald	Oak Ridge National Laboratory
Whitney	Armstrong	Argonne National Laboratory
Richard	Arthur	General Electric Research
Srinivasan	Arunajatesan	Sandia National Laboratories
Paul	Atzberger	University of California, Santa Barbara
Brian	Austin	Lawrence Berkeley National Laboratory
Ariful	Azad	Indiana University
Gyorgy	Babnigg	Argonne National Laboratory
Tyler	Backman	Lawrence Berkeley National Laboratory
Drew	Baden	Department of Energy, High Energy Physics

First Name	Last Name	Institution
David	Bader	New Jersey Institute of Technology
Jermon	Bafaty	Department of Energy, Artificial Intelligence and Technology Office
Zhe	Bai	Lawrence Berkeley National Laboratory
Ray	Bair	Argonne National Laboratory
Vamshi	Balanaga	Sandia National Laboratory/UC Berkeley
Prasanna	Balaprakash	Argonne National Laboratory
Jan	Balewski	Lawrence Berkeley National Laboratory
Mark	Bandstra	Lawrence Berkeley National Laboratory
Feng	Bao	Florida State University
David	Barajas-Solano	Pacific Northwest National Laboratory
Giuseppe	Barbalinardo	University of California, Davis
Deborah	Bard	Lawrence Berkeley National Laboratory
Jaydeep	Bardhan	GlaxoSmithKline
Ashley	Barker	Oak Ridge National Laboratory
Richard	Barnes	Lawrence Berkeley National Laboratory
Kipton	Barros	Los Alamos National Laboratory
Edward	Barry	Argonne National Laboratory
Robert	Bartolo	Transformational Liaisons (TRL), LLC
Bipul	Barua	Argonne National Laboratory
Jennifer	Bauer	National Energy Technology Laboratory
Alex	Bayen	University of California, Berkeley
Matthew	Becker	Argonne National Laboratory
Pete	Beckman	Argonne National Laboratory
Bo	Begole	AMD Research
James	Belak	Lawrence Livermore National Laboratory
Matt	Bement	Los Alamos National Laboratory
Douglas	Benjamin	Argonne National Laboratory
Janine	Bennett	Sandia National Laboratories
Russell	Bent	Los Alamos National Laboratory
Timothy	Berg	Sandia National Laboratories
Joshua	Bergerson	Argonne National Laboratory
Anne	Berres	Oak Ridge National Laboratory
Michael	Berube	Department of Energy
Julie	Bessac	Argonne National Laboratory
Wes	Bethel	Lawrence Berkeley National Laboratory
Budhu	Bhaduri	Oak Ridge National Laboratory
Wahid	Bhimji	Lawrence Berkeley National Laboratory
Debsihdu	Bhowmik	Oak Ridge National Laboratory
Sirui	Bi	Oak Ridge Institute for Science and Education
Tekin	Bicer	Argonne National Laboratory
Sandra	Biedron	Element Aero
Hassina	Bilheux	Oak Ridge National Laboratory
Jean	Bilheux	Oak Ridge National Laboratory
Jay Jay	Billings	Oak Ridge National Laboratory

First Name	Last Name	Institution
Adam	Bingston	Oak Ridge National Laboratory
Steve	Binkley	Department of Energy
Jens	Birkholzer	Lawrence Berkeley National Laboratory
Larry	Birnbaum	Northwestern University
Ayan	Biswas	Los Alamos National Laboratory
Laura	Biven	Department of Energy, Advanced Scientific Computing Research
Rocco	Blais	National Intelligence University
Arthur	Bland	Oak Ridge National Laboratory
Willem	Blokland	Oak Ridge National Laboratory
Josh	Bloom	Lawrence Berkeley National Laboratory
Swen	Boehm	Oak Ridge National Laboratory
Amber	Boehnlein	Jefferson Laboratory
John	Boger	Department of Energy
Dorian	Bohler	SLAC National Accelerator Laboratory
Trudy	Bolin	University of Wisconsin, Milwaukee
Lynn	Borkon	Frederick National Laboratory
Nikolay	Borodinov	Oak Ridge National Laboratory
Kristofer	Bouchard	Lawrence Berkeley National Laboratory
Charles	Bouman	Purdue University
Alison	Boyer	Oak Ridge National Laboratory
Mark	Boyer	Princeton Plasma Physics Laboratory
Selen	Bozkurt	Stanford University
Tom	Brady	Dell Technologies
Jim	Brandt	Sandia National Laboratories
Justin H. S.	Breaux	Argonne National Laboratory
Peer-Timo	Bremer	Lawrence Livermore National Laboratory
Thomas S.	Brettin	Argonne National Laboratory
Ron	Brightwell	Sandia National Laboratories
Michael	Brim	Oak Ridge National Laboratory
Grant	Bromhal	National Energy Technology Laboratory
David	Bross	Argonne National Laboratory
Ben	Brown	Department of Energy, Advanced Scientific Computing Research
David	Brown	Lawrence Berkeley National Laboratory
J. Ben	Brown	Lawrence Berkeley National Laboratory
Acacia	Brunett	Argonne National Laboratory
Mark	Buckner	Oak Ridge National Laboratory
Aydin	Buluc	Lawrence Berkeley National Laboratory
Keith	Burghardt	University of Southern California
Shawn	Burns	Sandia National Laboratories
Ralph	Butler	Argonne National Laboratory/Middle Tennessee State University
Suren	Byna	Lawrence Berkeley National Laboratory
John	Byrd	Argonne National Laboratory
Viveck	Cadambe	Pennsylvania State University

First Name	Last Name	Institution
Helen	Cademartori	Lawrence Berkeley National Laboratory
Hao	Cai	Argonne National Laboratory
Zhonghou	Cai	Argonne National Laboratory
Paolo	Calafiura	Lawrence Berkeley National Laboratory
Kelly	Callison	Information International Associates, Inc
John	Canik	Oak Ridge National Laboratory
Shane	Canon	Lawrence Berkeley National Laboratory
Yue	Cao	Argonne National Laboratory
Jian	Cao	Northwestern University
Adam	Cardinal-Stakenas	National Security Agency, Research
Suma	Cardwell	Sandia National Laboratories
Altaf	Carim	Department of Energy, High Energy Physics
Richard	Carlson	Department of Energy
Jonathan	Carter	Lawrence Berkeley National Laboratory
Emily	Casleton	Los Alamos National Laboratory
Vic	Castillo	Lawrence Livermore National Laboratory
Charlie	Catlett	Argonne National Laboratory
Christine	Chalk	Department of Energy
Maria	Chan	Argonne National Laboratory
Emory	Chan	Lawrence Berkeley National Laboratory
Cy	Chan	Lawrence Berkeley National Laboratory
Hau	Chan	University of Nebraska, Lincoln
Jin	Chang	California Institute of Technology
Shing	Chang	Kansas State University
Hang	Chang	Lawrence Berkeley National Laboratory
CS (Choongseok)	Chang	Princeton Plasma Physics Laboratory
Lali	Chatterjee	Department of Energy, High Energy Physics
Arghya	Chatterjee	Oak Ridge National Laboratory
Santanu	Chaudhuri	Argonne National Laboratory
Julio Jonas	Chaves Montero	Argonne National Laboratory
Saurabh	Chawdhary	Argonne National Laboratory
Weiyang	Chen	Argonne National Laboratory
Jinsong	Chen	Lawrence Berkeley National Laboratory
Barry	Chen	Lawrence Livermore National Laboratory
Wei	Chen	Northwestern University
Jieyang	Chen	Oak Ridge National Laboratory
Jacqueline	Chen	Sandia National Laboratories
Jian	Chen	The Ohio State University/Interactive Visual Computing Lab
Shunda	Chen	University of California, Davis
Alvin	Cheung	University of California, Berkeley
Andrew	Chien	Argonne National Laboratory
Taylor	Childers	Argonne National Laboratory

First Name	Last Name	Institution
Eric	Chisolm	Department of Energy, National Nuclear Security Administration
Jong Youl	Choi	Oak Ridge National Laboratory
Swati	Choudhary	Calysta
Alok	Choudhary	Northwestern University
Souma	Chowdhury	University at Buffalo
Marshall	Choy	SambaNova Systems
Hans	Christen	Oak Ridge National Laboratory
Blair	Christian	Oak Ridge National Laboratory
Giri	Chukkapalli	NVIDIA Corporation
Sudheer	Chunduri	Argonne National Laboratory
Michael	Churchill	Princeton Plasma Physics Laboratory
Jennifer	Clarke	University of Nebraska
Ian	Cloet	Argonne National Laboratory
Daniel	Clouse	Department of Defense
Ryan	Coffee	SLAC National Accelerator Laboratory
Susan	Coghlan	Argonne National Laboratory
Mark	Coletti	Oak Ridge National Laboratory
Jim	Collins	Argonne National Laboratory
William	Collins	Lawrence Berkeley National Laboratory
Scott	Collis	Argonne National Laboratory
Samuel	Collis	Sandia National Laboratories
Guojing	Cong	IBM Research
Emil	Constantinescu	Argonne National Laboratory
Simon	Corrodi	Argonne National Laboratory
Andrea	Cortis	Belmont Technology
Chip	Cotton	General Electric Research
Sarah	Cousineau	Oak Ridge National Laboratory
Stephen	Crago	University of Southern California, ISI
Claire	Cramer	Department of Energy
Valentino	Crespi	University of Southern California, ISI
Jody	Crisp	Oak Ridge Institute for Science and Education
Ethan	Crumlin	Lawrence Berkeley National Laboratory
Claire	Curry	Bloomberg
Matthew	Curry	Sandia National Laboratories
Larry	Curtiss	Argonne National Laboratory
Christine	Custis	NewPearl, Inc.
Christine	Cutillo	National Institutes of Health, NCATS
Eric	Cyr	Sandia National Laboratories
Ed	D’Azevedo	Oak Ridge National Laboratory
Paul	Dabbar	Department of Energy
Jamison	Daniel	Oak Ridge National Laboratory
Payel	Das	IBM Research
Debolina	Dasgupta	Argonne National Laboratory
Ganesh	Dasika	AMD Research

First Name	Last Name	Institution
Warren	Davis	Sandia National Laboratories
Marcus	Day	Lawrence Berkeley National Laboratory
Maarten	de Hoop	Rice University
Wibe	de Jong	Lawrence Berkeley National Laboratory
Cees	de Laat	Lawrence Berkeley National Laboratory
Sebastian	De Pascuale	Oak Ridge National Laboratory
David	Dean	Oak Ridge National Laboratory
Victor	Decaria	Oak Ridge National Laboratory
Gemechis	Degaga	Oak Ridge National Laboratory
Anthony	DeGennaro	Brookhaven National Laboratory
Paramvir	Dehal	Lawrence Berkeley National Laboratory
Payman	Dehghanian	George Washington University
Diego	del Castillo Negrete	Oak Ridge National Laboratory
Phillip	DeLeon	New Mexico State University
Marcel	Demarteau	Oak Ridge National Laboratory
James	Demmel	University of California, Berkeley
Patric	Den Hartog	Argonne National Laboratory
Anton	Dereventsov	Oak Ridge National Laboratory
Riccardo	Dettori	University of California, Davis
Sheng	Di	Argonne National Laboratory
Zichao Wendy	Di	Argonne National Laboratory
Alexa	Di Paolo	Bloomberg
Lori	Diachin	Lawrence Livermore National Laboratory
Jorge	Diaz Cruz	University of New Mexico\ SLAC
Emily	Dietrich	Argonne National Laboratory
Spiros	Dimolitsas	Georgetown University
Chao	Ding	Lawrence Berkeley National Laboratory
Nan	Ding	Lawrence Berkeley National Laboratory
Remi	Dingreville	Sandia National Laboratories
Stanley	Dodds	University of Hawaii/Institute for Astronomy
Emily	Donahue	Sandia National Laboratories
Sijia	Dong	Argonne National Laboratory
Ge	Dong	Princeton Plasma Physics Laboratory
Jack	Dongarra	University of Tennessee
Jana	Doppa	Washington State University
Max	Dornfest	Lawrence Berkeley National Laboratory
Sudip	Dosanjh	Lawrence Berkeley National Laboratory
Mathieu	Doucet	Oak Ridge National Laboratory
Ye	Duan	University of Missouri
Javier	Duarte	Fermi National Accelerator Laboratory
Nicolas	Dube	Hewlett Packard Enterprise
Vincent	Dumont	Lawrence Berkeley National Laboratory
Daniel	Dunlavy	Sandia National Laboratories
Soumya	Dutta	Los Alamos National Laboratory
Shyam	Dwaraknath	Lawrence Berkeley National Laboratory
Dipankar	Dwivedi	Lawrence Berkeley National Laboratory

First Name	Last Name	Institution
Carol	Eddy-Dilek	Savannah River National Laboratory
Romain	Egele	Argonne National Laboratory
Markus	Eisenbach	Oak Ridge National Laboratory
Muammar	El Khatib	Lawrence Berkeley National Laboratory
V. Daniel	Elvira	Fermi National Accelerator Laboratory
Wael	Elwasif	Oak Ridge National Laboratory
Murali	Emani	Argonne National Laboratory
Sujata	Emani	Department of Energy, BER
Eirik	Endeve	Oak Ridge National Laboratory
Christian	Engelmann	Oak Ridge National Laboratory
Sarah	Eno	University of Maryland
Peter	Ercius	Lawrence Berkeley National Laboratory
Ali	Erdemir	Argonne National Laboratory
Stephane	Ethier	Princeton Plasma Physics Laboratory
David	Etim	Department of Energy, National Nuclear Security Administration
Kate	Evans	Oak Ridge National Laboratory
Tom	Evans	Oak Ridge National Laboratory
VJ	Ewing	Oak Ridge National Laboratory
Farah	Fahim	Fermi National Accelerator Laboratory
Fariba	Fahroo	Air Force Office of Scientific Research
Chris	Fall	Department of Energy
George	Fann	Oak Ridge National Laboratory
Paolo	Faraboschi	Hewlett Packard Enterprise
Amro	Farid	Dartmouth College
Steven	Farrell	Lawrence Berkeley National Laboratory
Pooyan	Fazli	San Francisco State University
Tingzhou	Fei	Argonne National Laboratory
Frank	Felder	Rutgers University
Yan	Feng	Argonne National Laboratory
Wu	Feng	Virginia Tech
Phil	Ferguson	Oak Ridge National Laboratory
Nicola	Ferrier	Argonne National Laboratory
Emily	Fetter	Boston University
Hal	Finkel	Argonne National Laboratory
Nicole	Fisk	Cray, Inc.
Mary	Fitzpatrick	Argonne National Laboratory
Aaron	Fluitt	Argonne National Laboratory
Thomas	Flynn	Brookhaven National Laboratory
David	Fobes	Los Alamos National Laboratory
Fernanda	Foertter	NVIDIA Corporation
Ian	Foster	Argonne National Laboratory
Guillaume	Fouche	Bloomberg
Geoffrey	Fox	Indiana University
Kelly	Gaither	The University of Texas at Austin
Alexey	Galda	Argonne National Laboratory

First Name	Last Name	Institution
Alfredo	Galindo-Uribarri	Oak Ridge National Laboratory
Jack	Gallant	University of California, Berkeley
Yu	Gan	University of Alabama
Baskar	Ganapathysubramanian	Iowa State University
Rishi	Ganeriwala	Lawrence Livermore National Laboratory
Hector	Garcia Martin	Lawrence Berkeley National Laboratory
Marta	Garcia Martinez	Argonne National Laboratory
Arti	Garg	Cray, Inc.
Krishna	Garikipati	University of Michigan
Christopher	Garland	Argonne National Laboratory
Andrew	Gaspar	Los Alamos National Laboratory
Gerald	Geernaert	Department of Energy
R.Stuart	Geiger	University of California, Berkeley
Al	Geist	Oak Ridge National Laboratory
Ann	Gentile	Sandia National Laboratories
Cole	Gentry	Oak Ridge National Laboratory
Martina	Gerbino	Argonne National Laboratory
Tim	Germann	Los Alamos National Laboratory
Berk	Geveci	Kitware, Inc.
Mehran	Ghafari	University of Tennessee at Chattanooga
Devarshi	Ghoshal	Lawrence Berkeley National Laboratory
Erwin	Gianchandani	National Science Foundation
Tom	Gibbs	NVIDIA Corporation
Scott	Gibson	Oak Ridge National Laboratory
Michael	Giering	United Technologies/Pratt & Whitney
Roscoe	Giles	Boston University
Roberto	Gioiosa	Pacific Northwest National Laboratory
Shawn	Gleason	Oak Ridge National Laboratory
David	Gleich	Purdue University
Sergei	Gleyzer	University of Alabama/Fermilab
Jennifer	Glore	SambaNova Systems
Eric	Goodman	Sandia National Laboratories
Daniel	Gopman	National Institute of Standards and Technology
Ben	Gould	Dell EMC
Marco	Govoni	Argonne National Laboratory
Peter	Graf	National Renewable Energy Laboratory
Carlo	Graziani	Argonne National Laboratory
Emily	Greenspan	National Cancer Institute
Susan	Gregurick	National Institutes of Health
Annette	Greiner	Lawrence Berkeley National Laboratory
Michael	Grosskopf	Los Alamos National Laboratory
Allan	Grosvenor	Microsurgeonbot Inc.
Ray	Grout	National Renewable Energy Laboratory
Taylor	Groves	Lawrence Berkeley National Laboratory
Amy	Gryshuk	Lawrence Livermore National Laboratory

First Name	Last Name	Institution
Qiang	Guan	Kent State University/Los Alamos National Laboratory
Mamikon	Guillan	Sandia National Laboratories
Donna	Guillen	Idaho National Laboratory
Max	Gunzburger	Oak Ridge National Laboratory
Hanqi	Guo	Argonne National Laboratory
Haobo	Guo	University of Tennessee at Chattanooga
Chin	Guok	Lawrence Berkeley National Laboratory
Geetika	Gupta	NVIDIA Corporation
Hoshin	Gupta	University of Arizona
Dogan	Gursoy	Argonne National Laboratory
Tejas	Guruswamy	Argonne National Laboratory
Benjamin	Gutierrez-Garcia	Argonne National Laboratory
Salman	Habib	Argonne National Laboratory
James	Hack	Oak Ridge National Laboratory
Kawtar	Hafidi	Argonne National Laboratory
Aric	Hagberg	Los Alamos National Laboratory
Shima	Hajimirza	Texas A&M University
Mahantesh	Halappanavar	Pacific Northwest National Laboratory
Jason	Hales	Idaho National Laboratory
Charlotte	Haley	Argonne National Laboratory
Scot	Halverson	Los Alamos National Laboratory
Kathleen	Hamilton	Oak Ridge National Laboratory
Jeff	Hammond	Intel Corporation
Steve	Hammond	National Renewable Energy Laboratory
T. Yong	Han	Lawrence Livermore National Laboratory
Briana	Hanafin	Accenture
Marcus	Hanwell	Kitware, Inc.
Zhao	Hao	Lawrence Berkeley National Laboratory
Bruce	Hardy	Savannah River National Laboratory
Rachel	Harken	Oak Ridge National Laboratory
Kevin	Harms	Argonne National Laboratory
Peter	Harrington	Lawrence Berkeley National Laboratory
William	Hart	Sandia National Laboratories
Cory	Hauck	Oak Ridge National Laboratory
Nancy	Hayden	Sandia National Laboratories
Andrew	Hearin	Argonne National Laboratory
Sean	Hearne	Oak Ridge National Laboratory
Matt	Heavner	Los Alamos National Laboratory
Alexander	Heifetz	Argonne National Laboratory
Nils	Heinonen	Argonne National Laboratory
Olle	Heinonen	Argonne National Laboratory
Alan	Heirich	SLAC National Accelerator Laboratory
Katrin	Heitmann	Argonne National Laboratory
Barbara	Helland	Department of Energy, Office of Science
Bruce	Hendrickson	Lawrence Livermore National Laboratory

First Name	Last Name	Institution
Nicolas	Hengartner	Los Alamos National Laboratory
Marc	Henry de Frahan	National Renewable Energy Laboratory
Tina	Hernandez-Boussard	Stanford University
Michael	Heroux	Sandia National Laboratories
Kenneth	Herwig	Oak Ridge National Laboratory
Joel	Hestness	Cerebras Systems
Alexander	Hexemer	Lawrence Berkeley National Laboratory
Tony	Hey	SciML Group, Rutherford Appleton Lab, UK
Judith	Hill	Oak Ridge National Laboratory
Lindsey	Hillesheim	Cray, Inc.
Jacob	Hinkle	Oak Ridge National Laboratory
Jeffrey	Hittinger	Lawrence Livermore National Laboratory
Justin	Hnilo	Department of Energy
Phay	Ho	Argonne National Laboratory
Thuc	Hoang	Department of Energy, National Nuclear Security Administration
Andy	Hock	Cerebras Systems
Torsten	Hoefler	ETH Zurich
Forrest M.	Hoffman	Oak Ridge National Laboratory
Sabine	Hollatz	Stanford University
Brian	Homerding	Argonne National Laboratory
Vasant	Honavar	Pennsylvania State University
Tianzhen	Hong	Lawrence Berkeley National Laboratory
Walter	Hopkins	Argonne National Laboratory
Chet	Hopp	Lawrence Berkeley National Laboratory
Paul	Hovland	Argonne National Laboratory
Stephan	Hoyer	Google Research
Elizabeth	Hsu	National Cancer Institute
Lucy	Hsu	National Institutes of Health, NHLBI
Michael	Hu	Argonne National Laboratory
Rui	Hu	Argonne National Laboratory
Bin	Hu	Los Alamos National Laboratory
Xiang	Huang	Argonne National Laboratory
Yu	Huang	Argonne National Laboratory
Xiaobiao	Huang	SLAC National Accelerator Laboratory
Eliu	Huerta	University of Illinois at Urbana-Champaign
Ashley	Huff	Oak Ridge National Laboratory
David	Hughes	Oak Ridge National Laboratory
Travis	Humble	Oak Ridge National Laboratory
Sean	Hurley	California Polytechnic State University
Lorraine	Hwang	University of California, Davis
Hoon	Hwangbo	University of Tennessee
Costin	Iancu	Lawrence Berkeley National Laboratory
Khaled	Ibrahim	Lawrence Berkeley National Laboratory
Matthew	Igel	University of California, Davis
Gabriel	Ilevbare	Idaho National Laboratory

First Name	Last Name	Institution
Nwike	Iloeje	Argonne National Laboratory
Ilse C.F.	Ipsen	North Carolina State University
Ehsan Sabri	Islam	Argonne National Laboratory
Robert	Jackson	Argonne National Laboratory
Robert	Jacob	Argonne National Laboratory
Chris	Jacobsen	Argonne National Laboratory/Northwestern University
Dan	Jacobson	Oak Ridge National Laboratory
Anubhav	Jain	Lawrence Berkeley National Laboratory
Ralph	James	Savannah River National Laboratory
Kathy	Jang	University of California, Berkeley
Michael	Jarrett	Argonne National Laboratory
Cynthia	Jenks	Argonne National Laboratory
Elise	Jennings	Argonne National Laboratory
Vince	Jesaitis	Arm Inc
Shantenu	Jha	Brookhaven National Laboratory
Yi	Jiang	Argonne National Laboratory
Zhenhua	Jiang	University of Dayton Research Institute
Meng	Jiang	University of Notre Dame
Xiao-Yong	Jin	Argonne National Laboratory
Mingzhou	Jin	University of Tennessee
Marcin	Joachimiak	Lawrence Berkeley National Laboratory
Eugene	John	University of Texas at San Antonio
Fred	Johnson	Department of Energy, Retired
Travis	Johnston	Oak Ridge National Laboratory
Eric	Jonas	University of Chicago
Gregory	Jones	Oak Ridge National Laboratory
Katie	Jones	Oak Ridge National Laboratory
Scott	Jones	Oak Ridge National Laboratory
Terry	Jones	Oak Ridge National Laboratory
Doug	Joseph	Arm Inc
Renu	Joseph	Department of Energy
Earl	Joseph	Hyperion Research
Wayne	Joubert	Oak Ridge National Laboratory
Gary	Jung	Lawrence Berkeley National Laboratory
Andrew	Kail	Savannah River National Laboratory
Rajiv	Kalia	University of Southern California
Sergei	Kalinin	Oak Ridge National Laboratory
Mingon	Kang	University of Nevada, Las Vegas
Ramakrishnan	Kannan	Oak Ridge National Laboratory
Mahmut	Karakaya	University of Central Arkansas
Ulas	Karaoz	Lawrence Berkeley National Laboratory
Ian	Karlin	Lawrence Livermore National Laboratory
Alisha	Kasam-Griffith	Argonne National Laboratory
Karthik	Kashinath	Lawrence Berkeley National Laboratory
Aggelos	Katsaggelos	Northwestern University

First Name	Last Name	Institution
Kimberly	Kaufeld	Los Alamos National Laboratory
Brian	Kaul	Oak Ridge National Laboratory
Aditya	Kaushal	Bank of Montreal
Henry	Kautz	National Science Foundation, CISE
Trevor	Keenan	Lawrence Berkeley National Laboratory
Ken	Kemner	Argonne National Laboratory
Kelly	Kessler	Bloomberg
Rajkumar	Kettimuthu	Argonne National Laboratory
Foaad	Khosmood	California Polytechnic State University
Kathy	Kincade	Lawrence Berkeley National Laboratory
Ryan	King	National Renewable Energy Laboratory
Jeffery	Kinnison	Argonne National Laboratory/University of Notre Dame
Mariam	Kiran	Lawrence Berkeley National Laboratory
Uma	Klaassen	Oak Ridge National Laboratory
Hilda	Klasky	Oak Ridge National Laboratory
Scott	Klasky	Oak Ridge National Laboratory
Kerstin	Kleese van Dam	Brookhaven National Laboratory
Stanley	Klein	University of California, Berkeley
Tim	Kneafsey	Lawrence Berkeley National Laboratory
Christopher	Knight	Argonne National Laboratory
Katie	Knight	Oak Ridge National Laboratory
Ryan	Knox	Lawrence Berkeley National Laboratory
Tamara	Kolda	Sandia National Laboratories
Egemen	Kolemen	Princeton University
Kadidia	Konate	Lawrence Berkeley National Laboratory
Urs	Koster	Cerebras Systems
Rao	Kotamarthi	Argonne National Laboratory
Olivera	Kotevska	Oak Ridge National Laboratory
Doug	Kothe	Oak Ridge National Laboratory
John	Koudelka	Idaho National Laboratory
William	Kramer	University of Illinois/NCSA
James	Kress	Oak Ridge National Laboratory
Harinarayan	Krishnan	Lawrence Berkeley National Laboratory
Ralph	Kube	Princeton Plasma Physics Laboratory
Paul	Kuberry	Sandia National Laboratories
Michelle	Kuchera	Davidson College
Suhas	Kumar	Hewlett Packard Laboratory
Dinesh	Kumar	Lawrence Berkeley National Laboratory
Jitu	Kumar	Oak Ridge National Laboratory
Praveen	Kumar	University of Illinois
Vinod	Kumar	University of Texas at El Paso/Calysta Inc.-Menlo Park
Ana	Kupresanin	Lawrence Livermore National Laboratory
Tom	Kurfess	Oak Ridge National Laboratory
Teja	Kuruganti	Oak Ridge National Laboratory

First Name	Last Name	Institution
Joshua	Ladau	Lawrence Berkeley National Laboratory
Yue Shi	Lai	Lawrence Berkeley National Laboratory
M. Paul	Laiu	Oak Ridge National Laboratory
Matthew	Lanctot	Department of Energy, Office of Science
TJ	Lane	SLAC National Accelerator Laboratory
Michael	Lang	Los Alamos National Laboratory
James	Laros	Sandia National Laboratories
Jeffrey	Larson	Argonne National Laboratory
Randall	Laviolette	Department of Energy, Advanced Scientific Computing Research
Earl	Lawrence	Los Alamos National Laboratory
Craig	Lawrence	University of Maryland
Nam	Le	Johns Hopkins University Applied Physics Lab
Jacqueline	Le Moigne	NASA Earth Science Technology Office
Damien	Lebrun-Grandie	Oak Ridge National Laboratory
Timothy	Ledlow	Missile Defense Agency
Eungje	Lee	Argonne National Laboratory
Steven	Lee	Department of Energy, Advanced Scientific Computing Research
Victor	Lee	Intel Corporation
Seyong	Lee	Oak Ridge National Laboratory
Ti	Leggett	Argonne National Laboratory
Remi	Lehe	Lawrence Berkeley National Laboratory
Margaret	Lentz	Department of Energy, Artificial Intelligence and Technology Office
Vitus	Leung	Sandia National Laboratories
Dawn	Levy	Oak Ridge National Laboratory
Katherine	Lewis	Lawrence Livermore National Laboratory
Katie	Lewis	Lawrence Livermore National Laboratory
Sven	Leyffer	Argonne National Laboratory
Meimei	Li	Argonne National Laboratory
Ying	Li	Argonne National Laboratory
Sherry	Li	Lawrence Berkeley National Laboratory
Ying Wai	Li	Los Alamos National Laboratory
Zhaojian	Li	Michigan State University
Ang	Li	Pacific Northwest National Laboratory
Dong	Li	University of California, Merced
Bo	Li	University of Illinois at Urbana-Champaign
Dmitry	Liakh	Oak Ridge National Laboratory
Dong	Liang	University of Maryland Center for Environmental Science
Chen	Liao	Argonne National Laboratory
Sean	Liddick	Michigan State University, NSCL
Meifeng	Lin	Brookhaven National Laboratory
Yuewei	Lin	Brookhaven National Laboratory

First Name	Last Name	Institution
Youzuo	Lin	Los Alamos National Laboratory
Eric	Lin	National Institute of Standards and Technology
Guang	Lin	Purdue University
Zhihong	Lin	University of California, Irvine
Travis	Linderman	Innovation DuPage - NIU/COD
Robert	Link	Pacific Northwest National Laboratory
Yung	Liu	Argonne National Laboratory
Cong	Liu	Argonne National Laboratory
Zhengchun	Liu	Argonne National Laboratory
Miaoyuan	Liu	Fermi National Accelerator Laboratory
Yan	Liu	Oak Ridge National Laboratory
Frank	Liu	Oak Ridge National Laboratory/CSMD
Jing	Liu	Stanford University
Bill	Livezey	Microsoft Corporation
Li-Ta	Lo	Los Alamos National Laboratory
Jeremy	Logan	Oak Ridge National Laboratory
Wolfgang	Losert	University of Maryland, College Park
Pavel	Lougovski	Oak Ridge National Laboratory
Dan	Lu	Oak Ridge National Laboratory
Xiaobin	Lu	Oak Ridge National Laboratory
Zarija	Lukic	Lawrence Berkeley National Laboratory
Dalton	Lunga	Oak Ridge National Laboratory
Feng	Luo	Clemson University
Lixiang	Luo	IBM Research
Xuan	Luo	Lawrence Berkeley National Laboratory
Bethany	Lusch	Argonne National Laboratory
Piotr	Luszczek	University of Tennessee
Joseph	Lykken	Fermi National Accelerator Laboratory
Steven	Lyness	Cray, Inc.
Adam	Lyon	Fermi National Accelerator Laboratory
Charles	Macal	Argonne National Laboratory
Arthur Barney	Maccabe	Oak Ridge National Laboratory
Michael	MacNeil	Lawrence Berkeley National Laboratory
Siddharth	Maddali	Argonne National Laboratory
Ravi	Madduri	Argonne National Laboratory
Sandeep	Madireddy	Argonne National Laboratory
Ramana	Madupu	Department of Energy
Gina	Magnotti	Argonne National Laboratory
Ketan	Maheshwari	Oak Ridge National Laboratory
Michael	Mahoney	University of California, Berkeley
Michael	Majurski	National Institute of Standards and Technology
Nicholas	Malaya	Advanced Micro Devices Company
Carlos	Maltzahn	University of California, Santa Cruz
Andrea	Manning	Argonne National Laboratory

First Name	Last Name	Institution
Arun Kumar	Mannodi Kanakkithodi	Argonne National Laboratory
Jiafu	Mao	Oak Ridge National Laboratory
Don	March	Oak Ridge National Laboratory
Phil	Markham	Southern Company
David	Martin	Argonne National Laboratory
Victoria	Martin	Argonne National Laboratory
Daniel	Martin	Lawrence Berkeley National Laboratory
Mark	Martin	Oak Ridge National Laboratory
Carianne	Martinez	Sandia National Laboratories
Ghoncheh	Mashayekhi	University of Wisconsin, Milwaukee
Zachary	Matheson	Department of Energy, National Nuclear Security Administration
Michael	Matheson	Oak Ridge National Laboratory
Romit	Maulik	Argonne National Laboratory
Yury	Maximov	Los Alamos National Laboratory
Ed	May	Argonne National Laboratory
Jessica	Mazerik	National Institutes of Health
Matt	McConnell	Dell EMC
Dana	McCoskey	Water Power Tech Office
Dylan	McDowell	Idaho National Laboratory
Cynthia	McMurray	Lawrence Berkeley National Laboratory
Hugh	Medal	University of Tennessee
Shafigh	Mehraeen	University of Illinois at Chicago
Apurva	Mehta	National Accelerator Laboratory\ SLAC
Kshitij	Mehta	Oak Ridge National Laboratory
Veronica	Melesse Vergara	Oak Ridge National Laboratory
Matt	Menickelly	Argonne National Laboratory
Bronson	Messer	Oak Ridge National Laboratory
Zein-Eddine	Meziani	Argonne National Laboratory
Georgios	Michelogiannakis	Lawrence Berkeley National Laboratory
Anitescu	Mihai	Argonne National Laboratory
Mark	Miller	Lawrence Livermore National Laboratory
David	Miller	National Energy Technology Laboratory
Richard	Mills	Argonne National Laboratory
Ryan	Milner	Argonne National Laboratory
Michael	Minion	Lawrence Berkeley National Laboratory
Sandeep	Miryala	Fermi National Accelerator Laboratory
Konstantin	Mischaikow	Rutgers University
Umakant	Mishra	Argonne National Laboratory
Utkarsh	Mital	Lawrence Berkeley National Laboratory
John	Mitchell	Argonne National Laboratory
Julie	Mitchell	Oak Ridge National Laboratory
John	Mitchell	Sandia National Laboratories
Susan	Mniszewski	Los Alamos National Laboratory
Daniel	Moberg	Argonne National Laboratory
Bashir	Mohammed	Lawrence Berkeley National Laboratory

First Name	Last Name	Institution
Subhasish	Mohanty	Argonne National Laboratory
Linda	Mohanty	Dell EMC
William	Monday	Oak Ridge National Laboratory
Inder	Monga	Lawrence Berkeley National Laboratory
Laura	Monroe	Los Alamos National Laboratory
Luis	Montero	Argonne National Laboratory
Allison	Montroy	Department of Defense, Air Force Research Laboratory
Elisabeth (Lissa)	Moore	Los Alamos National Laboratory
Juston	Moore	Los Alamos National Laboratory
Shirley	Moore	Oak Ridge National Laboratory
Mark	Moraes	D. E. Shaw Research
Kenneth	Moreland	Sandia National Laboratories
Hannah	Morgan	Argonne National Laboratory
Dmitriy	Morozov	Lawrence Berkeley National Laboratory
James	Morris	Ames Laboratory
Juliane	Mueller	Lawrence Berkeley National Laboratory
Terrell	Mundhenk	Lawrence Livermore National Laboratory
Todd	Munson	Argonne National Laboratory
Robert	Murray	Microsoft Corporation
Mustafa	Mustafa	Lawrence Berkeley National Laboratory
Srideep	Musuvathy	Sandia National Laboratories
Balu	Nadiga	Los Alamos National Laboratory
Ambarish	Nag	National Renewable Energy Laboratory
Habib	Najm	Sandia National Laboratories
Aiichiro	Nakano	University of Southern California
Hai Ah	Nam	Los Alamos National Laboratory
Brad	Nance	Oak Ridge National Laboratory
Youssef	Nashed	Argonne National Laboratory
Thomas	Naughton	Oak Ridge National Laboratory
Gary	Navrotski	Argonne National Laboratory
Thomas	Ndousse-Fetter	Department of Energy
Kyle	Neal	Sandia National Laboratories
Grey	Nearing	University of Alabama
Benjamin	Nebgen	Los Alamos National Laboratory
Tommy	Nelson	Oak Ridge National Laboratory
Denise	Neudecker	Los Alamos National Laboratory
Michelle	Newcomer	Lawrence Berkeley National Laboratory
Justin	Newcomer	Sandia National Laboratories
Harvey	Newman	California Institute of Technology
Ben	Newton	Sandia National Laboratories
Esmond	Ng	Lawrence Berkeley National Laboratory
Brenda	Ng	Lawrence Livermore National Laboratory
Marcus	Nguyen	Argonne National Laboratory/University of Chicago
Jeffrey	Nichols	Oak Ridge National Laboratory

First Name	Last Name	Institution
Bogdan	Nicolae	Argonne National Laboratory
Marcus	Noack	Lawrence Berkeley National Laboratory
Jorge	Nocedal	Northwestern University
Eva	Nogales	Lawrence Berkeley National Laboratory
Brian	Nord	Fermi National Accelerator Laboratory
Peter	Nugent	Lawrence Berkeley National Laboratory
Hoot	O'Connor	My Math Cloud
Patrick	O'Leary	Kitware, Inc.
Daniel	O'Malley	Los Alamos National Laboratory
Aleksandr	Obabko	Argonne National Laboratory
Ceferino	Obcemea	National Cancer Institute
Ron	Oldfield	Sandia National Laboratories
Lenny	Oliker	Lawrence Berkeley National Laboratory
Kunle	Olukotun	SambaNova Systems
Olufemi	Omitaomu	Oak Ridge National Laboratory
Raymond	Osborn	Argonne National Laboratory
Jim	Ostrowski	University of Tennessee
Sarah	Owens	Argonne National Laboratory
John	Owens	University of California, Davis
Opeoluwa	Owoyele	Argonne National Laboratory
Diane	Oyen	Los Alamos National Laboratory
Ozgur	Ozmen	Oak Ridge National Laboratory
Aaron	Packman	Northwestern University/Argonne National Laboratory
David	Page	Oak Ridge National Laboratory
Pinaki	Pal	Argonne National Laboratory
Dhabaleswar K (DK)	Panda	The Ohio State University
Achalesh Kumar	Pandey	General Electric Research
Tara	Pandya	Oak Ridge National Laboratory
Theo	Papamarkou	Oak Ridge National Laboratory
Michael E.	Papka	Argonne National Laboratory
Vincent	Paquit	Oak Ridge National Laboratory
Gilchan	Park	Brookhaven National Laboratory
Ji Hwan	Park	Brookhaven National Laboratory
Yoonho	Park	IBM Research
Eun Jung	Park	Los Alamos National Laboratory
Byung Hoon	Park	Oak Ridge National Laboratory
Lynne	Parker	Office of Science and Technology Policy
Valerio	Pascucci	University of Utah
Gilberto	Pastorello	Lawrence Berkeley National Laboratory
Deep	Patel	Oak Ridge National Laboratory
Abani	Patra	Tufts University
Christina	Patricola	Lawrence Berkeley National Laboratory
Robert	Patton	Oak Ridge National Laboratory
Robert	Pavel	Los Alamos National Laboratory
Chuck	Pavloski	Pennsylvania State University

First Name	Last Name	Institution
Roger	Pawlowski	Sandia National Laboratories
Kevin	Pedro	Fermi National Accelerator Laboratory
Sean	Peisert	Lawrence Berkeley National Laboratory
Amra	Peles	Pacific Northwest National Laboratory
Swann	Perarnau	Argonne National Laboratory
Talita	Perciano	Lawrence Berkeley National Laboratory
Gabriel	Perdue	Fermi National Accelerator Laboratory
Mauro	Perego	Sandia National Laboratories
Kalyan	Perumalla	Oak Ridge National Laboratory
Nick	Peters	Oak Ridge National Laboratory
Norm	Peterson	Argonne National Laboratory
Matt	Peterson	Sandia National Laboratories
Armenak	Petrosyan	Oak Ridge National Laboratory
Charudatta	Phatak	Argonne National laboratory
Bobby	Philip	Los Alamos National Laboratory
Caleb	Phillips	National Renewable Energy Laboratory
Cynthia	Phillips	Sandia National Laboratories
Mary Ann	Piette	Lawrence Berkeley National Laboratory
Ali	Pinar	Sandia National Laboratories
Robinson	Pino	Department of Energy
Steve	Plimpton	Sandia National Laboratories
Matthew	Plumlee	Northwestern University
Norbert	Podhorszki	Oak Ridge National Laboratory
Raphael	Pooser	Oak Ridge National Laboratory
Alex	Pothen	Purdue University
Thomas	Potok	Oak Ridge National Laboratory
Carol	Pott	Lawrence Berkeley National Laboratory
Line	Pouchard	Brookhaven National Laboratory
Sarah	Powers	Oak Ridge National Laboratory
	Prabhat	Lawrence Berkeley National Laboratory
Thomas	Proffen	Oak Ridge National Laboratory
Andrey	Prokopenko	Oak Ridge National Laboratory
James	Proudfoot	Argonne National Laboratory
Fernanda	Psihas	Fermi National Accelerator Laboratory/The University of Texas at Austin
Dave	Pugmire	Oak Ridge National Laboratory
Laura	Pullum	Oak Ridge National Laboratory
Ji	Qiang	Lawrence Berkeley National Laboratory
Hong	Qin	University of Tennessee at Chattanooga
Judy	Qiu	Indiana University
Alejandro	Queiruga	Lawrence Berkeley National Laboratory
John	Quigley	Dell EMC
Jofrey	Quintanar	Argonne National Laboratory
Mihaela	Quirk	Department of Energy, National Nuclear Security Administration
Brian	Quiter	Lawrence Berkeley National Laboratory

First Name	Last Name	Institution
Sudarsan	Rachuri	Department of Energy
Maryam	Rahnemoonfar	University of Maryland, Baltimore County
Gulshan	Rai	Department of Energy, Office of Nuclear Physics
Pankaj	Rajak	Argonne National Laboratory
Siva	Rajamanickam	Sandia National Laboratories
Hridesh	Rajan	Ames Laboratory/Iowa State University
Vinay	Ramakrishnaiah	Los Alamos National Laboratory
Lavanya	Ramakrishnan	Lawrence Berkeley National Laboratory
Arvind	Ramanathan	Argonne National Laboratory
Jini	Ramprakash	Argonne National Laboratory
Pradeep	Ramuhalli	Oak Ridge National Laboratory
Huzefa	Rangwala	George Mason University
Vishwas	Rao	Argonne National Laboratory
Nageswara	Rao	Oak Ridge National Laboratory
William	Ratcliff	National Institute of Standards and Technology
Daniel	Ratner	SLAC National Accelerator Laboratory
Jaideep	Ray	Sandia National Laboratories
Justin	Reese	Lawrence Berkeley National Laboratory
Ernst	Rehm	Argonne National Laboratory
Yihui	Ren	Brookhaven National Laboratory
Viktor	Reshniak	Oak Ridge National Laboratory
Randal	Rheinheimer	Los Alamos National Laboratory
James	Ricci	Department of Energy, Advanced Scientific Computing Research
Daniel	Ricciuto	Oak Ridge National Laboratory
Jasmin	Richard	University of Rochester
Elias	Rigas	CCDC Army Research Laboratory
Hugo	Riggs	Florida International University
Todd	Ringler	Rep. Ben Ray Luján
Benjamin	Robbins	Cray, Inc.
Mike	Robinson	Department of Energy, Wind Energy Technology Office
Verónica	Rodríguez Tribaldos	Lawrence Berkeley National Laboratory
Dmitry	Romanov	Jefferson Laboratory
Elisa	Romero Romero	University of Tennessee
Mohammad	Roni	Idaho National Laboratory
Kelly	Rose	National Energy Technology Laboratory
Derek	Rose	Oak Ridge National Laboratory
Michael	Rosenfield	IBM Research
Elizabeth	Rosenthal	Oak Ridge National Laboratory
Robert	Ross	Argonne National Laboratory
Fred	Rothganger	Sandia National Laboratories
Lindsay	Roy	Savannah River National Laboratory
Ahmad	Rushdi	Sandia National Laboratories

First Name	Last Name	Institution
Thomas	Russell	Department of Energy, Basic Energy Sciences
Florin	Rusu	Lawrence Berkeley National Laboratory
Gary	Saavedra	Sandia National Laboratories
Ella	Saccon	National Cancer Institute
Sonia	Sachs	Department of Energy, Office of Science
Cosmin	Safta	Sandia National Laboratories
Alec	Sandy	Argonne National Laboratory
Ramanan	Sankaran	Oak Ridge National Laboratory
Daniel	Santiago	Argonne National Laboratory
Fadil	Santosa	University of Minnesota
Jibo	Sanyal	Oak Ridge National Laboratory
Vivek	Sarkar	Georgia Institute of Technology
Mina	Sartipi	University of Tennessee at Chattanooga
Arif	Sarwat	Florida International University
Bhima	Sastri	Office of Fossil Energy
Paul	Saxe	Virginia Tech, MolSSI
Michael	Schatz	Georgia Institute of Technology
Ben	Schiltz	Argonne National Laboratory
John	Schlueter	National Science Foundation
Martin	Schoenball	Lawrence Berkeley National Laboratory
Malachi	Schram	Pacific Northwest National Laboratory
Robert	Schreiber	Cerebras Systems
Katie	Schuman	Oak Ridge National Laboratory
Michelle	Schwalbe	National Academies of Sciences, Engineering, and Medicine
Ann	Schwartz Drobnis	Computing Community Consortium
Ariel	Schwartzman	SLAC National Accelerator Laboratory
Nicholas	Schwarz	Argonne National Laboratory
Mary	Scott	Lawrence Berkeley National Laboratory
Sudip	Seal	Oak Ridge National Laboratory
Pablo	Seleson	Oak Ridge National Laboratory
Uros	Seljak	Lawrence Berkeley National Laboratory
Daisy Flora	Selvaraj	University of North Dakota
Satyabrata	Sen	Oak Ridge National Laboratory
Koushik	Sen	University of California, Berkeley
Sergio	Servantez	Argonne National Laboratory/Northwestern University
Robert	Service	Science Magazine
Jamie	Sethian	Lawrence Berkeley National Laboratory
Bradley	Settlemyer	Los Alamos National Laboratory
Gökhan	Sever	Argonne National Laboratory
William	Severa	Sandia National Laboratories
Volkan	Sevim	Lawrence Berkeley National Laboratory
James	Sexton	IBM Research
Elizabeth	Sexton-Kennedy	Fermi National Accelerator Laboratory

First Name	Last Name	Institution
John	Shalf	Lawrence Berkeley National Laboratory
Hairong	Shang	Argonne National Laboratory
Arjun	Shankar	Oak Ridge National Laboratory
susmit	shannigrahi	Tennessee Technological University
Himanshu	Sharma	Argonne National Laboratory
Akshay	Sharma	Lawrence Berkeley National Laboratory
Karlie	Sharma	National Institutes of Health, NCATS
Emily	Shemon	Argonne National Laboratory
Chaopeng	Shen	Pennsylvania State University
Huanjie	Sheng	University of California, Berkeley
Wei	Shi	National Energy Technology Laboratory/ LRST/Battelle
Xiaoying	Shi	Oak Ridge National Laboratory
Xinghua	Shi	Temple University
Galen	Shipman	Los Alamos National Laboratory
Cyna	Shirazinejad	University of California, Berkeley
Shalki	Shrivastava	University of North Carolina at Chapel Hill, RENCI
Forrest	Shriver	Oak Ridge National Laboratory
Maulik	Shukla	Argonne National Laboratory
Christopher	Siefert	Sandia National Laboratories
Andrew	Siegel	Argonne National Laboratory
Horst	Simon	Lawrence Berkeley National Laboratory
Sean	Simoneau	Oak Ridge National Laboratory
Rajneesh	Singh	US Army Research Lab
Ganesh	Sivaraman	Argonne National Laboratory
Adam	Slagell	Lawrence Berkeley National Laboratory
Stuart	Slattery	Oak Ridge National Laboratory
Tess	Smidt	Lawrence Berkeley National Laboratory
Barry	Smith	Argonne National Laboratory
Jeff	Smith	Oak Ridge National Laboratory
Michael	Smith	Oak Ridge National Laboratory
David	Smith	University of Wisconsin, Madison
Oleg	Sobolev	Lawrence Berkeley National Laboratory
Lynda	Soderholm	Argonne National Laboratory
Sibendu	Som	Argonne National Laboratory
Suhas	Somnath	Oak Ridge National Laboratory
Siamak	Sorooshyari	University of California, Berkeley
Salvador	Sosa Guitron	University of New Mexico
Carlos	Soto	Brookhaven National Laboratory
Dale	Southard	Groq Inc.
Brian	Spears	Lawrence Livermore National Laboratory
Maria	Spiropulu	California Institute of Technology
William	Spotz	Department of Energy
Michael	Sprague	National Renewable Energy Laboratory
Sarat	Sreepathi	Oak Ridge National Laboratory

First Name	Last Name	Institution
Niranjan	Sridhar	Verily Life Sciences
Srilok	Srinivasan	Argonne National Laboratory
Jay	Srinivasan	Lawrence Berkeley National Laboratory
Gowri	Srinivasan	Los Alamos National Laboratory
Peter	St. John	National Renewable Energy Laboratory
Andrew	Stack	Oak Ridge National Laboratory
Marius	Stan	Argonne National Laboratory
Vitalii	Starchenko	Oak Ridge National Laboratory
Janice	Steckel	National Energy Technology Laboratory
Chad	Steed	Oak Ridge National Laboratory
Carl	Steefel	Lawrence Berkeley National Laboratory
Carolyn	Steele	Argonne National Laboratory
Rick	Stevens	Argonne National Laboratory
Jim	Stewart	Sandia National Laboratories
Panos	Stinis	Pacific Northwest National Laboratory
Miroslav	Stoyanov	Oak Ridge National Laboratory
Tjerk	Straatsma	Oak Ridge National Laboratory
David	Stracuzzi	Sandia National Laboratories
Stephen	Streiffer	Argonne National Laboratory
Frederick	Streitz	Department of Energy, HQ
Forrest	Striver	Oak Ridge National Laboratory
Erich	Strohmaier	Lawrence Berkeley National Laboratory
Jan	Strube	Pacific Northwest National Laboratory
Abby	Stylianou	Saint Louis University
Eric	Suchyta	Oak Ridge National Laboratory
Sreenivas	Sukumar	Cray, Inc.
Bobby G.	Sumpter	Oak Ridge National Laboratory
Yipeng	Sun	Argonne National Laboratory
Chengjun	Sun	Argonne National Laboratory
Zhao	Sun	Hampton University
Yu	Sun	Stony Brook University
Shivshankar	Sundaram	Lawrence Livermore National Laboratory
Ceren	Susut	Department of Energy, Office of Science
Kamlesh	Suthar	Argonne National Laboratory
Carolin	Sutter-Fella	Lawrence Berkeley National Laboratory
Amy	Swain	Department of Energy
Pieter	Swart	Los Alamos National Laboratory
Christine	Sweeney	Los Alamos National Laboratory
Laura	Swiler	Sandia National Laboratories
Madhava	Syamlal	Department of Energy
Adam	Szymanski	Argonne National Laboratory
Michael	Tamillow	NICO
Jifu	Tan	Northern Illinois University
Yu-Hang	Tang	Lawrence Berkeley National Laboratory
Deepti	Tanjore	Lawrence Berkeley National Laboratory
Alexandre	Tartakovsky	Pacific Northwest National Laboratory

First Name	Last Name	Institution
Marc	Taubenblatt	IBM Research
Michela	Taufer	University of Tennessee
Valerie	Taylor	Argonne National Laboratory
Aniket	Tekawade	Argonne National Laboratory
Jeremy	Templeton	Sandia National Laboratories
Chris	Tennant	Jefferson Laboratory
Alan	Tennant	Oak Ridge National Laboratory
Kazuhiro	Terao	SLAC National Accelerator Laboratory
Guilhem	Tesseyre	Google Research
Rajeev	Thakur	Argonne National Laboratory
Jayakar	Thangaraj	Fermi National Accelerator Laboratory
Nicholas	Thompson	Oak Ridge National Laboratory
Aidan	Thompson	Sandia National Laboratories
Suzy	Tichenor	Oak Ridge National Laboratory
Ken	Tobin	Oak Ridge National Laboratory
Peter	Tonner	National Institute of Standards and Technology
Roberto	Torelli	Argonne National Laboratory
Georgia	Tourassi	Oak Ridge National Laboratory
Nhan	Tran	Fermi National Accelerator Laboratory
Hoang	Tran	Oak Ridge National Laboratory
Nathaniel	Trask	Sandia National Laboratories
Charles	Tripp	National Renewable Energy Laboratory
Andrew	Tritt	Lawrence Berkeley National Laboratory
Aristeidis	Tsaris	Oak Ridge National Laboratory
Bill	Turenne	Department of Energy, Artificial Intelligence and Technology Office
John	Turner	Oak Ridge National Laboratory
Sean	Turner	Pacific Northwest National Laboratory
Victor	Udeowa	General Services Administration
Thomas	Uram	Argonne National Laboratory
Nathan	Urban	Los Alamos National Laboratory
Meltem	Urgun-Demirtas	Argonne National Laboratory
Ahmet	Uysal	Argonne National Laboratory
Brian	Van Essen	Lawrence Livermore National Laboratory
Peter	van Gemmeren	Argonne National Laboratory
William	Vanderlinde	Department of Energy, Advanced Scientific Computing Research
Dirk	VanEssendelft	National Energy Technology Laboratory
Charuleka	Varadharajan	Lawrence Berkeley National Laboratory
Laurie	Varma	Oak Ridge National Laboratory
Robert	Varner	Oak Ridge National Laboratory
Natalia	Vasileva	Cerebras Systems
Dilip	Vasudevan	Lawrence Berkeley National Laboratory
Ashish	Vaswani	Google Research
Sudharshan	Vazhkudai	Oak Ridge National Laboratory

First Name	Last Name	Institution
Carolyn	Vea Lauzon	Department of Energy, HQ
Singanallur	Venkatakrishnan	Oak Ridge National Laboratory
Becky	Verastegui	Oak Ridge National Laboratory
Matthew	Verber	University of North Carolina at Chapel Hill
Rafael	Vescovi	Argonne National Laboratory
Velimir	Vesselinov	Los Alamos National Laboratory
Jeffrey	Vetter	Oak Ridge National Laboratory
Michael	Vildibill	Hewlett Packard Enterprise
Venkatram	Vishwanath	Argonne National Laboratory
Lukas	Vlcek	University of Tennessee
Charlie	Vollmer	Sandia National Laboratories
James	von Oehsen	Rutgers University
Dave	Vorhaus	Schmidt Futures
Greg	Wagner	Northwestern University
Robert	Wagner	Oak Ridge National Laboratory
Haruko	Wainwright	Lawrence Berkeley National Laboratory
Jay	Walsh	Northwestern University
Matthew	Walter	Toyota Technological Institute at Chicago
Cheng	Wang	Argonne National Laboratory
Haoyu	Wang	Argonne National Laboratory
Jiali	Wang	Argonne National Laboratory
Jin	Wang	Argonne National Laboratory
Bin	Wang	Lawrence Berkeley National Laboratory
Zhe	Wang	Lawrence Berkeley National Laboratory
Dali	Wang	Oak Ridge National Laboratory
Lipeng	Wang	Oak Ridge National Laboratory
Felix	Wang	Sandia National Laboratories
Zhang	Wanni	Lawrence Berkeley National Laboratory
Karl	Warburton	Iowa State University
Logan	Ward	Argonne National Laboratory
Sharlene	Weatherwax	Department of Energy, Biological and Environmental Research
Rosina	Weber	Drexel University
Gunther	Weber	Lawrence Berkeley National Laboratory
Clayton	Webster	Oak Ridge National Laboratory
Michael	Wehner	Lawrence Berkeley National Laboratory
Xishuo	Wei	University of California, Irvine
Patricia	Weikersheimer	Argonne National Laboratory
Jack	Wells	Oak Ridge National Laboratory
Haiden	Wen	Argonne National Laboratory
Torre	Wenaus	Brookhaven National Laboratory
Gerry	White	Federal Emergency Management Agency
Julia	White	Oak Ridge National Laboratory
Stephen	Whitelam	Lawrence Berkeley National Laboratory
Eric	Whiting	Idaho National Laboratory
Justin	Whitt	Oak Ridge National Laboratory

First Name	Last Name	Institution
Patrick	Widener	Sandia National Laboratories
Stefan	Wild	Argonne National Laboratory
George	Wilkie	Princeton Plasma Physics Laboratory
Sean	Wilkinson	Oak Ridge National Laboratory
Timothy	Williams	Argonne National Laboratory
Samuel	Williams	Lawrence Berkeley National Laboratory
Dan	Wilmot	Department of Energy, Artificial Intelligence and Technology Office
Peter	Winter	Argonne National Laboratory
Robert	Wisniewski	Intel Corporation
Laura	Wolf	Argonne National Laboratory
Matthew	Wolf	Oak Ridge National Laboratory
Michael	Wolf	Sandia National Laboratories
Phillip	Wolfram	Los Alamos National Laboratory
Gayle	Woloschak	Northwestern University
David	Womble	Oak Ridge National Laboratory
Geoff	Womeldorff	Los Alamos National Laboratory
Justin	Worrilow	Microsoft Corporation
Justin	Wozniak	Argonne National Laboratory
Nicholas	Wright	Lawrence Berkeley National Laboratory
Xuli	Wu	Argonne National Laboratory
Xingfu	Wu	Argonne National Laboratory/University of Chicago
Kesheng (John)	Wu	Lawrence Berkeley National Laboratory
Wei	Wu	Los Alamos National Laboratory
Sau Lan	Wu	University of Wisconsin, Madison
Margie	Wylie	Lawrence Berkeley National Laboratory
Max	Wyman	Argonne National Laboratory
Hai	Xiao	Clemson University
Lianghua	Xiong	Argonne National Laboratory
Yilun	Xu	Lawrence Berkeley National Laboratory
Zexuan	Xu	Lawrence Berkeley National Laboratory
Xueqiao	Xu	Lawrence Livermore National Laboratory
Min	Xu	Oak Ridge National Laboratory
Wenwei	Xu	Pacific Northwest National Laboratory
Yexiang	Xue	Purdue University
Chunhua	Yan	National Cancer Institute
Da	Yan	University of Alabama at Birmingham
Chao	Yang	Lawrence Berkeley National Laboratory
Da	Yang	Lawrence Berkeley National Laboratory
Zechun	Yang	Missile Defense Agency
Lexie	Yang	Oak Ridge National Laboratory
Qian	Yang	University of Connecticut
Ke-Thia	Yao	University of Southern California
Katherine	Yelick	Lawrence Berkeley National Laboratory
Orcun	Yildiz	Argonne National Laboratory

First Name	Last Name	Institution
Junqi	Yin	Oak Ridge National Laboratory
Shinjae	Yoo	Brookhaven National Laboratory
Kazutomo	Yoshii	Argonne National Laboratory
Linda	Young	Argonne National Laboratory/University of Chicago
Stanley	Young	National Renewable Energy Laboratory
Steven	Young	Oak Ridge National Laboratory
Andrew	Younge	Sandia National Laboratories
Shiqi	Yu	Argonne National Laboratory
Dantong	Yu	New Jersey Institute of Technology
Rose	Yu	Northeastern University
Thomas	Zacharia	Oak Ridge National Laboratory
Federico	Zahariev	Ames Laboratory
Nestor	Zaluzec	Argonne National Laboratory
Michael	Zarnstorff	Princeton Plasma Physics Laboratory
Piotr	Zarzycki	Lawrence Berkeley National Laboratory
Liat	Zavodivker	Lawrence Berkeley National Laboratory
Zuotao	Zeng	Argonne National Laboratory
Ruijie	Zeng	Utah State University
Hong	Zhang	Argonne National Laboratory
Xiaoyi	Zhang	Argonne National Laboratory
Yuepeng	Zhang	Argonne National Laboratory
Xiangyu	Zhang	National Renewable Energy Laboratory
Guannan	Zhang	Oak Ridge National Laboratory
Jiaxin	Zhang	Oak Ridge National Laboratory
Ying	Zhang	University of Rhode Island
Zhao	Zhang	University of Texas, TACC
Emma	Zhao	Argonne National Laboratory
Liang	Zhao	George Mason University
Huihuo	Zheng	Argonne National Laboratory
Zhi	Zheng	University of Wisconsin, Milwaukee
Mingxia	Zhou	Argonne National Laboratory
Maxim	Ziatdinov	Oak Ridge National Laboratory
Sue	Zillman	Argonne National Laboratory
Tarek	Zohdi	Lawrence Berkeley National Laboratory
Xiaobing	Zuo	Argonne National Laboratory
Petrus	Zwart	Lawrence Berkeley National Laboratory
Matthias	Zwicker	University of Maryland, College Park

AD. Abbreviations and Terminology

Abbreviations	Terminology
3D	three-dimensional
AGN	active galactic nucleus
AI	artificial intelligence
ALCF	Argonne Leadership Computing Facility
ALS	Advanced Light Source
AMIGA	All Modular Industry Growth Assessment
AMR	adaptive mesh refinement
ANNs	artificial neural networks
AOGCM	Atmosphere-ocean general circulation model
API	application programming interface
APS	appearance potential spectroscopy, Advanced Photon Source
Argonne	Argonne National Laboratory
ARM	atmospheric radiation monitoring
ARM	Atmospheric Radiation Measurement Climate Research Facility
ASCR	Advanced Scientific Computing Research
ASDEX-UG	Axially Symmetric Diverter Experiment Upgrade
BBH	binary black hole
Berkeley Lab	Lawrence Berkeley National Laboratory
BES	Basic Energy Sciences
BESAC	Basic Energy Sciences Advisory Committee
BG	Blue Gene
BHNS	black hole and neutron star
BNS	binary neutron star
CAF	Co-Array Fortran
CF	climate and forest
CGE	computable general equilibrium
CGRO	Compton Gamma-Ray Observatory
CMOS	complementary metal-oxide-semiconductor
CMS	Compact Muon Solenoid
CNNs	convolutional neural networks
CPU	central processing unit
CRISPR	clustered regularly interspaced short palindromic repeats
DAE	differential algebraic equation
DARPA	Defense Advanced Research Projects Agency
DAS	distributed acoustic sensing
DBA	design basis accident
DETF	Dark Energy Task Force
DFT	density functional theory
DL	deep learning
DLA	deep learning accelerator
DNN	deep neural network
DOE	United States Department of Energy
DVM	dynamic vegetation model

Abbreviations	Terminology
E3	Simulation and Modeling at the Exascale for Energy and the Environment
EAST	Experimental Advanced Superconducting Tokamak
ECoG	electrocorticography
EIC	Electron-Ion Collider
ELM	edge-localized mode
EMF	Energy Modeling Forum
EMSL	Environmental Molecular Sciences Laboratory
EOS	equation of state
ESGF	Earth System Grid Federation
ESM	Earth System Model
ESnet	Energy Sciences Network
ESS-DIVE	Environmental System Science Data Infrastructure for a Virtual Ecosystem
EVLA	Enhanced Very Large Array
EXIST	Energetic X-ray Imaging Survey Telescope
FAIR	findable, accessible, interoperable, reusable
FES	Fusion Energy Sciences
FFT	fast Fourier transform
flops	floating point operations per second
fMRI	functional magnetic resonance imaging
FPGA	field programmable gate array
FRIB	Facility for Rare Isotope Beams
FUSE	Far Ultraviolet Spectroscopic Explorer
GAN	generative adversarial network
Gbps	gigabits per second
GIS	geographic information system
GK	gyrokinetic
GRETA	Gamma-Ray Energy Tracking Array
GLAST	Gamma-ray Large Area Space Telescope
GMT	Giant Magellan Telescope
GNEP	Global Nuclear Energy Partnership
GPU	graphics processing unit
GRB	gamma-ray burst
GTC	Gyrokinetic Toroidal Code
HCCI	homogeneous charge compression ignition
HEP	high energy physics
HPC	high-performance computing
HPN	high-performance network
IEEE	Institute of Electrical and Electronics Engineers
I/O	Input/output
IOP	input/output processor
IoT	Internet of Things
Jefferson Lab	Thomas Jefferson National Accelerator Facility
JET	Joint European Torus
JUMP	Joint University Microelectronics Program
KBase	Systems Biology Knowledge Base
LAN	local area network

Abbreviations	Terminology
LBNL	Lawrence Berkeley National Laboratory
LCF	Leadership Computing Facility
LCLS-II	second-generation Linac Coherent Light Source
LSTM	long short-term memory
MD	molecular dynamic (simulations)
ML	machine learning
MPI	message passing interface
NERSC	National Energy Research Scientific Computing Center
NGEEs	Next-Generation Ecosystem Experiments
NIPS	Conference on Neural Information Processing Systems
NMDC	National Microbiome Data Collaborative
OLCF	Oak Ridge Leadership Computing Facility
ORNL	Oak Ridge National Laboratory
PLD	pulse laser deposition
QCD	quantum chromodynamics
QIS	quantum information sciences
RF	radio frequency
RHIC	Relativistic Heavy Ion Collider
RL	reinforcement learning
ROSM	reduced order surrogate model
SNS	Spallation Neutron Source
SoC	system-on-chip
SRF	superconducting radiofrequency
TB	terabyte
TPU	tensor processing unit
UHPC	ultra-high performance concrete
UQ	uncertainty quantification
WAN	wide area network

This page intentionally blank.

AE. References

01. Chemistry, Materials, and Nanoscience

1. Riordan, M. & Hoddeson, L., Crystal Fire: The Invention of the Transistor and the Birth of the Information Age, W. W. Norton & Company, 1998.
2. Sze, S. M., *Physics of Semiconductor Devices,* 2nd Edition, John Wiley and Sons, New York, 1981.
3. Shockley, W., Electrons and Holes in Semiconductors: With Applications to Transistor Electronics, D. Van Nostrand Company, Inc., 1950.
4. Fuechsle, M. et al., A single-atom transistor. *Nat. Nanotechnol.* **7**, 242–246 (2012).
5. Sumpter, B. G., Vasudevan, R. K., Potok, T., Kalinin, S. V., A bridge for accelerating materials design*. npj Comp. Mat.* **1**: 15008 (2015). DOI: 10.1038/npjcompumats. 2015.8
6. Kalinin, S. V., Sumpter, B. G., & Archibald, R. K., Big-deep-smart data in imaging for guiding materials design. *Nat. Mater.* **14**, 973–980 (2015).
7. M. Ziatdinov, et al., "Building and exploring libraries of atomic defects in graphene: Scanning transmission electron and scanning tunneling microscopy study," Sci. Adv. 5:eaaw8989 (2019). DOI: 10.1126/sciadv.aaw8989.

02. Earth and Environmental Sciences

1. Basu, S., Kumbier, K., Brown, J. B., & Yu, B. (2018). Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences*, *115*(8), 1943-1948.
2. Baydin, A. G., Shao, L., Bhimji, W., Heinrich, L., Meadows, L., Liu, J., & Ma, M. (2019). Etalumis: Bringing Probabilistic Programming to Scientific Simulators at Scale. *arXiv preprint arXiv:1907.03382*.
3. Bergen, K. J., Johnson, P. A., Maarten, V., & Beroza, G. C. (2019). Machine learning for data-driven discovery in solid Earth geoscience. *Science*, *363*(6433), eaau0323.
4. Bolton, Thomas, and Laure Zanna. "Applications of deep learning to ocean data inference and subgrid parameterization." *Journal of Advances in Modeling Earth Systems* 11, no. 1 (2019): 376-399.
5. Brantley, S. L. (2018) Shale Network Database, Consortium for Universities for the Advancement of Hydrologic Sciences, Inc. (CUAHSI). DOI: 10.4211/his-data-shalenetwork
6. Brenowitz, N. D., & Bretherton, C. S. (2018). Prognostic validation of a neural network unified physics parameterization. *Geophysical Research Letters*, 45, 6289-6298. https://doi.org/10.1029/2018GL07851
7. Cherukara, M. J., Nashed, Y. S. G., & Harder, R. J. Real-time coherent diffraction inversion using deep generative networks (2018). Scientific reports **8**(1), 165230.
8. Collins, W. & P. Tissot. An artificial neural network model to predict thunderstorms within 400 km^2 South Texas domain, *Meteorological Applications* 22, no. 3 (2015): 650-665.
9. Deng, J. et al.. Correlative 3D x-ray fluorescence and ptychographic tomography of frozen-hydrated green algae (2018), *Sci. Adv.***4**(11) eaau4548(1-10).
10. Flinchum, B. A., et al. Critical Zone Structure Under a Granite Ridge Inferred From Drilling and Three-Dimensional Seismic Refraction Data. (2018) J. Geophys. Res.: Earth Surf. 123 (6), 1317-1343.

11. Godinho, J. R. A., Gehrke, K. M., Stack, A. G., & Lee, P. D. (2016) The dynamic nature of crystal growth in pores. *Sci. Rep.,* 6:33086. DOI: 10.1038/srep33086

12. Hengl, T., et al. (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PLoS one*, *12*(2), e0169748.

13. Krasnopolsky, V., Nadiga, S., Mehra, A., Bayler, E., & Behringer, D. (2016). Neural networks technique for filling gaps in satellite measurements: Application to ocean color observations. *Computational Intelligence and Neuroscience* (2016): 29.

14. Kumar, J., Mills, R. T., Hoffman, F. M., & Hargrove, W. W. (2011). Parallel k-means clustering for quantitative ecoregion delineation using large data sets. *Procedia Computer Science*, *4*, 1602-1611.

15. Kurth, T. et al. (2018) Exascale deep learning for climate analytics. In *Proceedings of the International Conference for High Perfor- mance Computing, Networking, Storage, and Analysis*, pp. 51. IEEE Press,.

16. Laanait, N., He, Q., Borisevich, & A. Y. Reconstruction of 3-D Atomic Distortions from Electron Microscopy with Deep Learning. *arXiv, [cond-mat.mtrl-sci]*arXiv:1902.06876v1 19 Feb 2019

17. Liu, Y., Sun, W., & Durlofsky, L. J. (2019). A deep-learning-based geological parameter-ization for history matching complex models. *Mathematical Geosciences*, *51*(6), 725-766.

18. Li, Z., et al. (2016) Searching for anomalous methane in shallow groundwater near shale gas wells. *J. Contam. Hydrol.* 195, 23-30. DOI: 10.1016/j.jconhyd.2016.10.005

19. Lin, H.W., Tegmark, M., & Rolnick, D. Why Does Deep and Cheap Learning Work So Well? *J Stat Phys* (2017) 168: 1223.

20. Ling, F. T., et al. (2018) Nanospectroscopy Captures Nanoscale Compositional Zonation in Barite Solid Solutions. *Sci. Reports,*8:13041. DOI:10.1038/s41598-018-31335-3

21. Nogueira, K., Penatti, O. A., & dos Santos, J. A. (2017). Towards better exploiting convolutional neural networks for remote sensing scene classification. *Pattern Recognition*, *61*, 539-556.

22. O'Gorman, P. A., & Dwyer, J. G. (2018). Using machine learning to parameterize moist convection: Potential for modeling of climate, climate change, and extreme events. *Journal of Advances in Modeling Earth Systems*, 10, 2548–2563. https://doi.org/10.1029/2018MS001351

23. Rasp, Stephan, Michael S. Pritchard, and Pierre Gentine. "Deep learning to represent subgrid processes in climate models." *Proceedings of the National Academy of Sciences* 115, no. 39 (2018): 9684-9689.

24. Reichstein, M., et al. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195.

25. Scher, S. (2018). Toward Data-Driven Weather and Climate Forecasting: Approximating a Simple General Circulation Model With Deep Learning. *Geophysical Research Letters*, *45*(22), 12-616.

26. Schneider, T., Lan, S., Stuart, A., & Teixeira, J. (2017). Earth system modeling 2.0: A blueprint for models that learn from observations and targeted high-resolution simulations. *Geophysical Research Letters*, 44, 12,396-12,417. https://doi.org/10.1002/2017GL076101

27. Tartakovsky, M., C. Ortiz Marrero, P. Perdikaris, G. D. Tartakovsky, and D. Barajas-Solano, "Learning Parameters and Constitutive Relation- ships with Physics Informed Deep Neural Networks," arXiv e-prints, p. arXiv:1808.03398, Aug 2018.

28. Varadharajan et al., “Launching an Accessible Archive of Environmental Data,” Eos, vol. 100. 2019.

29. Wang, J., Balaprakash, P., and Kotamarthi, R. (2019). Fast domain-aware neural network emulation of a planetary boundary layer parameterization in a numerical weather forecast model, Geosci. Model Dev., 12, 4261–4274, https://doi.org/10.5194/gmd-12-4261-2019.

30. Zachara, J., et al. (2016) Internal Domains of Natural Porous Media Revealed: Critical Locations for Transport, Storage, and Chemical Reaction. *Environ. Sci. Technol.* 50, 2811-2829 DOI: 10.1021/acs.est.5b05015

31. Hoffman, F. M., et al. (2017. International Land Model Benchmarking (ILAMB) 2016 Workshop Report, Technical Report DOE/SC-0186, U.S. Department of Energy, Office of Science, Germantown, Maryland, USA, doi:10.2172/1330803.

32. https://www.ncdc.noaa.gov/billions/

33. http://www.energy.gov/downloads/usenergy-sector-vulnerabilities-climate-change-andextreme-weather

34. Zarzycki, P. Towards understanding of Reactive Interfaces in Geological CO2 Sequestration, RIGECO, ERC-2015-CoG Proposal 682274, September 2015.

03. Biology and Life Sciences

1. Garcia, B. J. et al. Phytobiome and Transcriptional Adaptation of *Populus deltoides* to Acute Progressive Drought and Cyclic Drought. *Phytobiomes Journal*. (2018) **2**(4), 249-60.

2. Bouchard K. E., et al. Union of Intersections (UoI) for Interpretable Data Driven Discovery and Prediction. *Advances in Neural Information Processing System*. (2017) **30**:1078-86.

3. Lawson C. E., et al. Common principles and best practices for engineering microbiomes. *Nat Rev Microbiol.* 2019. Epub 2019/09/25. doi: 10.1038/s41579-019-0255-9. PubMed PMID: 31548653.

4. Chmiela, S., et al. Towards exact molecular dynamics simulations with machine-learned force fields. *Nat. Commun.* **9**. 3887 (2018).

5. Murdoch, W. J. et al., Interpretable machine learning: definitions, methods, and applications. arXiv preprint. 2019.

6. Harnessing the Power of Data in Health. Stanford Medicine Health Trends Report. 2017.

7. Paddon, C. J., et al., High-level semi-synthetic production of the potent antimalarial artemisinin. *Nature* **496** 528-532 (25 April 2013).

8. Anderson, J. C., et al. Environmentally controlled invasion of cancer cells by engineered bacteria. *J. Mol. Biol.* 355(4):619-27 (27 Jan 2006).

9. Gardner, T. S. Synthetic biology: from hype to impact. *Trends Biotechnol*. **31**(3):123-5 (2013 Mar).

10. Gambhir, S. S., et al. Toward achieving precision health. *Sci. Transl. Med.* **10**(430) (28 Feb 2018).

11. Blaser, M. J., et al. Toward a predictive understanding of Earth’s microbiomes to address 21st century challenges. *Am. Soc. Microbiol.* (2016) doi: 10.1128/mBio.00714-16.

12. Allegretti, M., et al. Horizontal membrane-intrinsic α-helices in the stator a-subunit of an F-type ATP synthase. *Nature* **521**, 237-240 (14 May 2015).

13. Hermes, M., et al. Mid-IR hyperspectral imaging for label-free histopathology and cytology. J. Optics **20**(2) (24 Jan 2018).

14. Carbonell, P., T. Radivojevic and H. G. Martin. Opportunities at the Intersection of Synthetic Biology, Machine Learning, and Automation. *ACS Synth. Biol.* 2019, 8, 7, 1474-1477 (19 July 2019).

15. Census of Agriculture: Summary and State Data. United States Department of Agriculture. 2007.

16. Lal, R. Soil carbon sequestration to mitigate climate change. *Geoderma* **123**(1-2):1-22 (Nov 2004).

17. Hood, L. and L. Rowen. The Human Genome Project: big science transforms biology and medicine. *Genome Med*. **5**(9):79 (13 Sep 2013).

04. High Energy Physics

1. Rosner, J., et al., Planning the Future of US Particle Physics, *arXiv:1401.6075*

2. Cavuoti, S., et al., Machine-learning-based photometric redshifts for galaxies of the ESO Kilo-Degree Survey data release 2, *MNRAS 452, 3100 (2015).*

3. Kremer, J., et al., Big Universe, Big Data: Machine Learning and Image Analysis for Astronomy, *IEEE Intelligent Systems* 32, 16 (2017).

4. Higson, E., Handley, W., Hobson, M., and Lasenby, A., Bayesian sparse reconstruction: a brute force approach to astronomical imaging and machine learning, *MNRAS 483, 4828 (2019).*

5. Lanusse, F., et al., CMU DeepLens: deep learning for automatic image-based galaxy-galaxy strong lens finding, *MNRAS 473, 3895 (2018).*

6. Krause, E. and Eifler, T., CosmoLike – Cosmological Likelihood Analyses for Photometric Galaxy Surveys, *MNRAS, 470, 2100 (2017).*

7. Heitmann, K. et al., Cosmic Calibration, *Astrophys. J., 646, L1 (2006).*

8. Albertsson, K., et al., Machine Learning in High Energy Physics Community White Paper, *arXiv:1807.02876*

9. Radovic, A., et al., Machine learning at the energy and intensity frontiers of particle physics, *Nature 560, 41 (2018).*

10. Ilten, P., Williams, M., Yang, Y., Event generator tuning using bayesian optimization, *JINST 12.04 (2017).*

11. Albrect, J., HEP Community White Paper on Software trigger and event reconstruction, *arXiv: 1802.08638*

12. Collins, J. H., et al., Extending the Bump Hunt with Machine Learning, *arXiv: 1902.02634*

13. Ball, R.D., et al., Parton distributions for the LHC Run II, *JHEP 04, 40 (2015).*

14. Aurisano, A., et al., A Convolutional Neural Network Neutrino Event Classifier, *JINST 11.09 (2016).*

15. Acciarri, R., et al., Convolutional neural networks applied to neutrino events in a liquid argon time projection chamber, *JINST, 12.03 (2017).*

16. Akiyama, K., et al. (Event Horizon Telescope Collaboration), First M87 Event Horizon Telescope Results. I. The Shadow of the Supermassive Black Hole, *Astrophys. J. 875, L1 (2019).*

05. Nuclear Physics

1. Lee, I. Y., Gamma-ray tracking detectors. *Nucl. Instrum. Meth. A* **422**, 1-3 (1999), 195-200.

2. Deleplanque, M. A. et al., GRETA: utilizing new concepts in gamma-ray detection. *Nucl. Instrum. Meth. A* **430** 2-3 (1999), 292-310.

3. Goodfellow, I. et al., Generative Adversarial Nets. *Adv. Neural Inf. Process. Syst.* **27**, (2014) 2672–2680.

4. Gao, Y., J. Chen, T. Robertazzi, and K. A. Brown. Reinforcement learning based schemes to manage client activities in large distributed control systems. *Phys. Rev. Accel. Beams* **22**, 014601, January 2019.
5. Negoita, G. A., et al., Deep learning: Extrapolation tool for ab initio nuclear theory. *Phys. Rev. C* **99** (Oct. 2019).
6. Maaten, Laurens van der, and Geoffrey Hinton. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, Nov (2008): 2579-2605.
7. Dugger, M., et al., "A study of decays to strange final states with GlueX in Hall D using components of the BaBar DIRC," arXiv:1408.0215 [physics.ins-det].
8. Lai, Y. S., arXiv:1810.00835.
9. Ferrario, P. et al., "Demonstration of the event identification capabilities of the NEXT-White detector," arXiv:1905.13141 [physics.ins-det], accepted to JHEP 2019.
10. Solopova, A. D., et al., SRF Cavity Fault Classification Using Machine Learning at CEBAF. *Proc. 10th Int. Particle Accelerator Conf. (IPAC'19)*, Melbourne, Australia, May 2019, pp. 1167-1170.
11. Carpenter, A., et al., "Initial Implementation of a Machine Learning System for SRF Cavity Fault Classification at CEBAF." 17th Int. Conf. on Accelerator and Large Experimental Physics Control Systems (ICALEPCS'19), New York, NY, USA, Oct. 2019, paper WEPHA025.

06. Fusion

1. Gribov, Y., et al., "ITER Physics Basis," *Nuclear Fusion*, **47** (2007).
2. *Report of the Workshop on Advancing Fusion with Machine Learning April 30 – May 2, 2019.* https://science.osti.gov/-/media/fes/pdf/workshop-reports/FES_ASCR_Machine_Learning_Report.pdf
3. Baltz, E. A., et al., "Achievement of Sustained Net Plasma Heating in a Fusion Experiment with the Optometrist Algorithm," *Nature Scientific Reports*, **7** (2017). doi:10.1038/s41598-017-06645-7
4. Bock, A., et al., "Advanced Tokamak Investigations in Full-Tungsten ASDEX Upgrade," *Physics of Plasmas*, **25 (**2018).
5. Bonoli, P. T., et al., "Lower Hybrid Current Drive Experiments on Alcator C-Mod: Comparison with Theory and Simulation," *Physics of Plasmas*, **15** (2008).
6. Boyer, M. D., Kaye, S., Erickson, K. "Real-Time Capable Modeling of Neutral Beam Injection on NSTX-U Using Neural Networks," *Nuclear Fusion*, **59** (2019).
7. Cannas, B., Cau, F., Fanni, A., Sonato, P., Zedda, M.K., and JET-EFDA Contributors, "Automatic Disruption Classification at JET: Comparison of Different Pattern Recognition Techniques," *Nuclear Fusion*, **46** (2006).
8. Maingi, R., et al., "Summary of the FESAC Transformative Enabling Capabilities Panel Report," *Fusion Science and Technology,* **75** (2019).
9. Giruzzi, G., et al., "Physics and Operation Oriented Activities in Preparation of the JT-60SA Tokamak Exploitation," *Nuclear Fusion*, **57** (2017).
10. Gopalaswamy, V., et al., "Tripled Yield in Direct-Drive Laser Fusion through Statistical Modelling," *Nature*, **565** (2019).
11. Hill, D.N., et al., "DIII-D Research Towards Resolving Key Issues for ITER and Steady State Tokamaks," *Nuclear Fusion*, **53** (2013).
12. Kates-Harbeck, J., Svyatkovskiy, A., Tang, W., "Predicting Disruptive Instabilities in Controlled Fusion Plasmas Through Deep Learning," *Nature*, **568** (2019).

13. Li, J., et al, “A Long-Pulse High Confinement Plasma Regime in the Experimental Advanced Superconducting Tokamak,” *Nature Physics,* **9** (2013).

14. Meneghini, O., et al., “Self-Consistent Core-Pedestal Transport Simulations With Neural Network Accelerated Models,” *Nuclear Fusion*, **57** (2017).

15. Montes, K. J., et al., “Machine Learning for Disruption Warning on Alcator C-Mod, DIII-D, and EAST,” *Nuclear Fusion*, **59** (2019).

16. Rea, C., et al., “Disruption Prediction Investigations using machine learning tools on DIII-D and Alcator C-Mod,” *Plasma Physics and Controlled Fusion*, **60** (2018).

17. Rebut, P-H., “The Joint European Torus (JET),” European Physical Journal, 43 (2018).

18. Baker, N., et al. *Workshop Report on Basic Research Needs for Scientific Machine Learning: Core Technologies for Artificial Intelligence*. doi:10.2172/1478744 (2019).

19. Smith, R. C., “Uncertainty Quantification: Theory, Implementation, and Applications,” SIAM, Philadelphia (2014)

20. Windsor, C. G., Pautasso, G., Tichmann, C., Buttery, R. J., Hender, T. C., JET EFDA Contributors and the ASDEX-UG team, “A Cross-Tokamak Neural Network Disruption Predictor for the JET and ASDEX Upgrade Tokamaks,” *Nuclear Fusion*, **45** (2005).

21. Wroblewski, D., Jahns, G. L., Leuer, J. A., “Tokamak Disruption Alarm Based on a Neural Network Model of the High-Beta Limit,” *Nuclear Fusion*, **37** (1997).

07. Engineering and Manufacturing

1. Zistl, S. “The Future of Manufacturing: Prototype Robot Solves Problems without Programming,” *Seimens.com Global Website.*

2. Microsoft UK Enterprise Team. “Better, faster, more efficient: AI meets manufacturing.” *Microsoft Industry Blog – United Kingdom* (6 June, 2018).

3. “Airbus: Reimagining the future of air travel.” *Autodesk Website*.

4. *U.S. National Committee on Theoretical and Applied Mechanics, Board on International Scientific Organizations, Policy and Global Affairs, and National Academies of Sciences, Engineering, and Medicine, Predictive Theoretical and Computational Approaches for Additive Manufacturing: Proceedings of a Workshop.* Washington, D.C.: National Academies Press, 2016. DOI: 10.17226/23646.

5. *Board on Mathematical Sciences and Analytics, National Materials and Manufacturing Board, Division on Engineering and Physical Sciences, and National Academies of Sciences, Engineering, and Medicine, Data-Driven Modeling for Additive Manufacturing of Metals: Proceedings of a Workshop.* Washington, D.C.: National Academies Press, 2019. DOI: 10.17226/25481.

6. Bonawitz, K., et al., *Practical Secure Aggregation for Privacy-Preserving Machine Learning.* Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. 1175-1191. Oct 30-Nov 3, Dallas, TX, 2017.

7. Kasiviswanathan, S. P., et al. What Can We Learn Privately? *The 49th Annual IEEE Symposium on Foundations of Computer Science.* 531-540. Oct. 25-28, Philadelphia, PA (2008).

8. Balde, C. P., et al. *The Global E-waste Monitor 2017: Quantities, Flows, and Resources* (Bonn, Geneva, and Vienna: United Nations University, International Telecommunication Union, and International Solid Waste Association, 2017).

9. Ellen MacArthur Foundation, Circular Consumer Electronics: An Initial Exploration, 2018.

08. Smart Energy Infrastructure

1. Kwasinski, F., Andrade, M. J. Castro-Sitiriche and E. O'Neill-Carrillo, "Hurricane Maria Effects on Puerto Rico Electric Power Infrastructure," *IEEE Power and Energy Technology Systems Journal*, **6**, 85-94 (2019). doi: 10.1109/JPETS.2019.2900293
2. U.S.-Canada Power System Outage Task Force, *Final Report on the August 14, 2003 Blackout in the United States and Canada: Causes and Recommendations*, April 2004.
3. IEEE guide for electric power distribution reliability indices, IEEE Std 1366-2012.
4. U.S. Department of Energy website, "Confronting the Duck Curve: How to Address Over-Generation of Solar Energy," (2017).
5. National Academies of Sciences, Engineering, and Medicine 2017. Enhancing the Resilience of the Nation's Electricity System. Washington, DC: The National Academies Press. doi.org/10.17226/24836.
6. Hong, T., Chen, Y., Lee, S.H., Piette, M.A. "CityBES: A Web-based Platform to Support City-Scale Building Energy Efficiency," *Urban Computing* (2016).
7. Chen, Y., Hong, T., Piette, M.A. "Automatic Generation and Simulation of Urban Building Energy Models Based on City Datasets for City-Scale Building Retrofit Analysis," *Applied Energy* (2017).
8. U.S. Department of Energy, "Smart Grid System Report," (2018).
9. Hong, T., et al. "Ten questions on urban building energy modeling," *Building and Environment* (2019).
10. U.S. Department of Energy, Grid Interactive Efficient Buildings https://www.energy.gov/eere/buildings/grid-interactive-efficient-buildings
11. U.S. Department of Energy, Energy Efficient Mobility Systems https://www.energy.gov/eere/vehicles/energy-efficient-mobility-systems

09. AI for Computer Science

1. Ibrahim, A., Elfadel, M., Boning, D., Li, X. (Ed.), Machine Learning in VLSI Computer-Aided Design, Springer International Publishing, 2018.
2. Toigo, J., AI for Storage Management Gets Real. *Tech Target* (2019). https://www.google.com/amp/s/searchstorage.techtarget.com/opinion/AI-for-storage-management-gets-real%3famp=1
3. 2nd International Workshop on AI-assisted Design for Architecture https://eecs.oregonstate.edu/aidarc/index.php/program/
4. Bavishi, R., Lemieux, C., Fox, R., Sen, K., & Stoica, I., AutoPandas: Neural-Backed Generators for Program Synthesis. *Proceedings of the ACM on Programming Languages*, OOPSLA'19, October 2019.
5. Ansel, J., Kamil, et al., Opentuner: An extensible framework for program autotuning, *Proceedings of the 23rd International Conference on Parallel Architectures and Compilation*, 303–316. ACM, 2014.
6. Balaprakash, P., et al., Autotuning in High-performance Computing Applications, *Proceedings of the IEEE*, 1–16, 2018.
7. Tiwari, A., Chen, C., Chame, J., Hall, M., & Hollingsworth, J., A Scalable Auto-tuning Framework for Compiler Optimization, *Proceedings of the 2009 IEEE International Symposium on Parallel & Distributed Processing*, 1-12, 2009.

8. Thiagarajan, J. J., et al., Bootstrapping Parameter Space Exploration for Fast Tuning, *Proceedings of the 2018 International Conference on Supercomputing*, 385–395, November 2018.

9. Marathe, A., et al. Performance Modeling Under Resource Constraints Using Deep Transfer Learning, *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis* (SC17), 31, 2017.

10. Behzad, B., et al., Taming Parallel I/O Complexity with Auto-tuning, *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis* (SC13), 68, 2013.

11. Lin, X., Wang, Y., & Pedram, M., A Reinforcement Learning-based Power Management Framework for Green Computing Data Centers, 2016 IEEE International Conference on Cloud Engineering (IC2E), 2016.

12. Kalyan, A., et al., Neural-guided Deductive Search for Real-time Program Synthesis from Examples, The Sixth International Conference on Learning Representations (ICLR 2018), 2018.

13. Rao, N. S. V., Sen, S., Liu, Z., Kettimuthu, R., & Foster, I., Learning Concave-convex Profiles of Data Transport Over Dedicated Connections, Machine Learning for Networking, Springer-Verlag, 2019.

14. Cai, J, et al., Making Neural Programming Architectures Generalize Via Recursion, The Fifth International Conference on Learning Representations (ICLR 2017), 2017.

15. Sid-Lakhdar, W., Mahmoudi Aznaveh, Mohsen, M.A., Li, X., & Demmel, J., Multitask and Transfer Learning for Autotuning Exascale Applications, submitted August 2019.

16. Hoos, H.H., Programming by Optimization. *Communications of the ACM* **55**, 70–80 (2012). DOI: https://doi.org/10.1145/2076450.2076469

17. Berry, M., et al., *Machine Learning and Understanding for Intelligent Extreme Scale Scientific Computing and Discovery*, technical report, DOE ASCR Workshop Report, 2015.

18. Luan, S., Yang, D., Barnaby, C., Sen, K., & Chandra, S., Aroma: Code Recommendation via Structural Code Search,*, Proceedings of the ACM on Programming Languages* (OOPSLA'19), October 2019.

19. Cambronero, J., Li, H., Kim, S., Sen, K., & Chandra, S., When Deep Learning Met Code Search*, Industry Track of 27th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (ESEC/FSE'19), ACM, 964–974, August 2019.

20. Tramèr, F., et al. Ensemble adversarial training: Attacks and defenses. arXiv preprint arXiv:1705.07204 (2017).

21. ASCR Cybersecurity for Scientific Computing Integrity - Research Pathways and Ideas Workshop https://escholarship.org/content/qt5j00n7h2/qt5j00n7h2.pdf

22. Krishnan, S., J. Wang, E. Wu, M. J. Franklin, and K. Goldberg. 2016. ActiveClean: Interactive Data Cleaning for Statistical Modeling. *Proc. VLDB Endow*. **9**, 948–959 (2016).

23. Tarski, A.. *Logic, Semantics, Metamathematics: Papers from 1923 to 1938*, Oxford University Press, 1956.

24. Rao, N. S. V. On undecidability aspects of resilient computations and implications to Exascale, Resilience 2014: Seventh Workshop on Resiliency in High Performance Computing with Clouds, Grids, and Clusters, 2014.

25. Vapnik, V. N. *Statistical Learning Theory*. John-Wiley and Sons, New York, New York, 1998.

26. Cohen, F. B. “Computational aspects of computer virus,” *Computer & Security*, **8**, 325–344, 1989.

27. Rao, N. S. V., Reister, D. B., Barhen, J. Information Fusion Methods Based on Physical Laws*, IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**, 66–77 (2005).

28. Rao, N. S. V., et al. Multi-Modal Sensor Fusion for Reactor Power-Level Estimation: Thermal, EM, Acoustic. Nuclear Security Applications Research & Development Program Review Meeting, 2019.

29. Ben-David, S., Hrubes, P., Moran, S., Shpilka A., and Yehudayoff, A. *Learnability Can Be Undecidable*, Nature, 2019.

10. AI Foundations and Open Problems

1. Thomas, N., et al. Tensor Field Networks: Rotation- and Translation-Equivariant Neural Networks for 3D Point Clouds. Arxiv Preprint, arXiv:1802.08219, 2018.

2. Jordan, M. I., Artificial Intelligence: The Revolution Hasn’t Happened Yet. *Harvard Data Science Review* (2019). doi:10.1162/99608f92.f06c6e61.

3. Baker, N., et al. *Workshop Report on Basic Research Needs for Scientific Machine Learning: Core Technologies for Artificial Intelligence*, 2019. doi:10.2172/1478744

4. Arridge, S., Maass, P., Öktem, O., and Schönlieb, C. Solving Inverse Problems Using Data-Driven Models. *Acta Numerica,* **28**, 1-174. doi:10.1017/S0962492919000059.

5. Kondor, R., Trivedi, S. *On the Generalization of Equivariance and Convolution in Neural Networks to the Action of Compact Groups*. Proceedings of the 35th International Conference on Machine Learning, PMLR 80:2747–2755, 2018.

6. Goodfellow, I., et al. *Generative Adversarial Nets. Advances in Neural Information Processing Systems*, 2014.

7. Chen, X., et al. Infogan: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. *Advances in Neural Information Processing Systems,* 2172–2180 (2016).

8. Arjovsky, M., Chintala, S. and Bottou, L. *Wasserstein Generative Adversarial Networks.* International Conference on Machine Learning (pp. 214–223, 2017).

9. Paganini, M., de Oliveira, L. and Nachman, B. CaloGAN: Simulating 3D High Energy Particle Showers in Multilayer Electromagnetic Calorimeters with Generative Adversarial Networks*. Physical Review D* **97**: 014021 (2018).

10. Zhang, K., Zuo, W., Chen, Y., Meng, D. and Zhang, L. Beyond a Gaussian denoiser: Residual Learning of Deep CNN for Image Denoising. *IEEE Transactions on Image Processing,* **26:** 3142–3155 (2017).

11. He, K., Zhang, X., Ren, S. and Sun, J. Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 770–778, 2016.

12. Bottou, L., Curtis, F.E. and Nocedal, J. Optimization Methods for Large-Scale Machine Learning. *Siam Review,* **60:** 223–311 (2018).

13. LeCun, Y.A., Bottou, L., Orr, G.B. and Müller, K.R. *Efficient Backprop. Neural Networks: Tricks of the Trade*, Springer, Berlin, Heidelberg, 2012.

14. Sutskever, I., Martens, J., Dahl, G. and Hinton, G. On the Importance of Initialization and Momentum in Deep Learning. *International Conference on Machine Learning,* 1139–1147 (2013).

15. Duchi, J., Hazan, E., and Singer, Y. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 2121–2159 (2011).

16. National Research Council.– *Frontiers in Massive Data Analysis* Washington, DC: The National Academies Press. 2013. https://doi.org/10.17226/18374.

17. Mustafa, M., et al. CosmoGAN: Creating High-Fidelity Weak Lensing Convergence Maps Using Generative Adversarial Networks, *Computational Astrophysics and Cosmology* **6**, (2019).

18. Wu, J. L., et al. *Enforcing Statistical Constraints Generative Adversarial Networks for Modeling Chaotic Dynamical Systems*, Cornell University, 2019. https://arxiv.org/abs/1905.06841.

19. Raissi, M., Perdikaris, P., Karniadakis, *G. E. Physics Informed Deep Learning (Part I): Data-Driven Solutions of Nonlinear Partial Differential Equations*. https://arxiv.org/abs/1711.10561.

20. Yang, L., et al. Highly Scalable, Physics-Informed GANs for Learning Solutions of Stochastic PDEs (SC'19 Deep Learning on Supercomputers Workshop).

21. Weiler, M., Geiger, M., Welling, M., Boomsma, W. and Cohen, T. *3D Steerable CNNS: Learning Rotationally Equivariant Features in Volumetric Data.* Advances in Neural Information Processing Systems, 10381–10392, 2018.

22. T. D. Bui, S. Ravi and V. Ramavijjala, *Neural Graph Learning: Training Neural Networks Using Graphs*, Proceedings of 11th ACM International Conference on Web Search and Data Mining, 2018.

23. R. L. Murphy, B. Srinivasan, V. Rao and B. Ribeiro, Relational Pooling for Graph Representations, Arxiv:1903.02541, 2019.

24. K. Xu, W. Hu, J. Leskovec and S. Jegelka, How Powerful are Graph Neural Networks? ArXiv:1810.00826v3, 2019.

25. Tschannen, M., Bachem, O. and Lucic, M., 2018. Recent Advances in Autoencoder-Based Representation Learning. arXiv preprint arXiv:1812.05069.

26. Ben-David, S., Hrubeš, P., Moran, S. et al. Learnability Can be Undecidable*. Nat Mach Intell* **1:** 44–48 (2019). doi:10.1038/s42256-018-0002-3

27. Tshitoyan, V., et al. Unsupervised Word Embeddings Capture Latent Knowledge from Materials Science Literature. *Nature* **571**, 95–98 (2019). doi:10.1038/s41586-019-1335-8

28. Swain, M. C. and Cole, J. M., 2016. ChemDataExtractor: a Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *Journal of Chemical Information and Modeling*, **56**: 1894–1904 (2016).

29. Clark, P., et al., 2019. From 'F' to 'A' on the NY Regents Science Exams: An Overview of the Aristo Project. arXiv preprint arXiv:1909.01958.

11. Software Environments and Software Research

1. Baker, N., et al. Workshop Report on Basic Research Needs for Scientific Machine Learning: Core Technologies for Artificial Intelligence, DOE Office of Science Technical Report, 2019.

2. Gil, Y. & Selman, B., A 20-Year Community Roadmap for Artificial Intelligence Research in the US, 2019.

12. Data Life Cycle and Infrastructure

1. Wilkinson, M. D. et al. The FAIR Guiding Principles for Scientific Data Management and Stewardship (*Sci. Dat.* 3, 2016).

2. Blaiszik, B. et al., A Data Ecosystem to Support Machine Learning in Materials Science (*MRS Commun.*, 2019).
3. Himanen, L., Geurts, A., Foster, A. S., Rinke, P. Data-Driven Materials Science: Status, Challenges, and Perspectives (*Adv. Sci.*, 2019).
4. Arkin, A. P., et al. KBase: The United States Department of Energy Systems Biology Knowledgebase (*Nat. Biotechnol.* 36, 7, 2018).
5. Williams, D. N., et al. The Earth System Grid: Enabling Access to Multimodel Climate Simulation Data (*Bull. Am. Meteorol. Soc.* 90, 2, 195-206, 2009).
6. Stokes, G. M., & Schwartz, S. The Atmospheric Radiation Measurement Program (*Bull. Am. Meteorol. Soc.* 75, 7, 1201–1222, 1994).
7. Weber, G. H., Ophus, C., & Ramakrishnan, L. Automated Labeling of Electron Microscopy Images Using Deep Learning (*Proc. IEEE/ACM Mach. Learn. in HPC Environ.*, 26–36, 2018).
8. Biven, L., Office of Science Data for AI Roundtable: Presentation to ASCAC (http://bit.ly/2QWyTbr, 2019).
9. Aspuru-Guzik, A., & Persson, K. Materials Acceleration Platform: Accelerating Advanced Energy Materials Discovery by Integrating High-Throughput Methods and Artificial Intelligence (http://nrs.harvard.edu/urn-3:HUL.InstRepos:35164974, 2018).
10. Carbonell, P. Radivojevic, T., & García Martín, H. Opportunities at the Intersection of Synthetic Biology, Machine Learning, and Automation (*ACS Synth. Biol.* 8, 1474–1477, 2019).
11. Blaiszik, B., Charting ML Publications in Science (https://github.com/blaiszik/ml_publication_charts, 2019).
12. Chard, K., et al. The Modern Research Data Portal: A Design Pattern for Networked, Data-Intensive Science (*Peer J. Comput. Sci.* 4 e144, 2018).

13. Hardware Architectures

1. Chien, A. Computer Architecture: Disruption from Above, *Commun. ACM* **61**, 9, 2018.
2. Wu, C., et al., Machine Learning at Facebook: Understanding Inference at the Edge, Proceedings of the 2019 IEEE International Symposium on High Performance Computer Architecture (HPCA), 331–44 (https://doi.org/10.1109/HPCA.2019.00048, 2019).
3. LeCun, Y. Deep Learning Hardware: Past, Present, and Future, Proceedings of the 2019 IEEE International Solid-State Circuits Conference (ISSCC), 12–19 (https://doi.org/10.1109/ISSCC.2019.8662396, 2019).
4. Jouppi, N. P., et al., In-Datacenter Performance Analysis of a Tensor Processing Unit, SIGARCH Comput. Archit. News **45,** 2,1–12 (https://doi.org/10.1145/3140659.3080246, 2017).
5. Vetter, J. S., et al., Extreme Heterogeneity 2018 – Productive Computational Science in the Era of Extreme Heterogeneity: Report for DOE ASCR Workshop on Extreme Heterogeneity, USDOE Office of Science (https://www.osti.gov/servlets/purl/1473756, https://doi.org/10.2172/1473756, 2018).
6. AAAS Science Magazine, How Researchers are Teaching AI to Learn Like a Child, May 24, 2018.
7. Strubell, E., Ganesh, A., and McCallum, A. Energy and Policy Considerations for Deep Learning in NLP. (https://arxiv.org/abs/1906.02243)
8. Silver, D., et al., Mastering the game of Go without human knowledge, *Nature* **550**, 354, (https://doi.org/10.1038/nature24270, 2017).

9. Torrejon, J., et al., Neuromorphic computing with nanoscale spintronic oscillators, *Nature* **547**, 428, (https://doi.org/10.1038/nature23011, 2017).

10. Basic Research Needs for Microelectronics (Brochure), USDOE Office of Science (https://www.osti.gov/servlets/purl/15457, 72, 2018).

14. AI for Imaging

1. Liu, S., Leemann, S. C., Hexemer, A., Marcus, M. A., Melton, C. N., Nishimura, H., Sun, C. 2019. Demonstration of machine learning-based model-independent stabilization of source properties in synchrotron light sources. *PRL*, in press.

2. Kaira, C. S., et al. Automated correlative segmentation of large transmission x-ray microscopy (TXM) tomograms using deep learning. *Mater. Charact.* 142, 203–210 (2018).

3. Pelt, D. M., & Sethian J. A. A mixed-scale dense convolutional neural network for image analysis. *PNAS*, 115 (2) 254–259 (2018).

4. Chang, M.C., et al. Accelerating neutron scattering data collection and experiments using AI deep super-resolution learning (arXiv:1904.08450, 2019).

5. Samarakoon, A. N., et al. Machine learning assisted insight into spin ice $Dy_2Ti_2O_7$ (arXiv:1906.11275, 2019).

6. U.S. Department of Energy. Report from the Basic Energy Sciences Advisory Committee. Challenges at the frontiers of matter and energy: transformative opportunities for discovery science (2015).

7. U.S. Department of Energy. Report from the Biological and Environmental Research Advisory Committee. Grand Challenges for Biological and Environmental Research: Progress and Future Vision; A Report from the Biological and Environmental Research Advisory Committee, DOE/SC–0190, BERAC Subcommittee on Grand Research Challenges for Biological and Environmental Research (science.osti.gov/~/media/ber/berac/pdf/Reports/BERAC-2017-Grand-Challenges-Report.pdf, 2017).

8. The Advanced Photon Source Strategic Plan: Enabling frontier science in the national interest (2018).

9. ALS-U: Solving Scientific Challenges with Coherent Soft X-Rays (2017).

10. Noack, M. M., et al. A Kriging-Based Approach to Autonomous Experimentation with Applications to X-Ray Scattering, *Sci. Rep.* **9**, 11809 (2019).

11. Yang, X., et al. "Low-Dose X-Ray Tomography through a Deep Convolutional Neural Network." *Sci. Rep.* **8**, 2575 (2018).

15. AI at the Edge

1. "Edge Computing: Vision and Challenges," June 9, 2016 (https://ieeexplore.ieee.org/document/7488250, accessed October 11, 2019).

2. "Edge-centric Computing–DOIs," September 30, 2015, http://doi.org/10.1145/2831347.2831354, accessed October 11, 2019.

3. "Patterned Probes for High Precision 4D-STEM Bragg Measurements," July 11, 2019, https://arxiv.org/abs/1907.05504, accessed October 11, 2019.

4. "Making the Invisible Visible: New Sensor Network Reveals Telltale Patterns in Neighborhood Air Quality," July 22, 2019, https://newscenter.lbl.gov/2019/07/22/new-sensor-network-neighborhood-air-quality/, accessed October 11, 2019.

5. "Waggle: An open sensor platform for edge computing," https://ieeexplore.ieee.org/abstract/document/7808975/, accessed October 11, 2019.

6. "Array of things: a scientific research instrument in the public way," April 18, 2017, https://dl.acm.org/citation.cfm?id=3063771, accessed October 11, 2019.

7. “Argonne supports grid advances through pioneering energy storage and sensor research,” March 18, 2019, https://www.anl.gov/es/article/argonne-supports-grid-advances-through-pioneering-energy-storage-and-sensor-research, accessed October 11, 2019.
8. “Edge TPU–Google Cloud,” https://cloud.google.com/edge-tpu/, accessed October 11, 2019.
9. “Intel Movidius, an Intel Company,” https://www.movidius.com/, accessed October 11, 2019.
10. “Brain-inspired Chip–IBM Research,” http://www.research.ibm.com/articles/brain-chip.shtml, accessed October 11, 2019.
11. “AR1K–The Smart Farm Research Consortium,” https://ar1k.org/, accessed October 11, 2019.
12. “AR1K: Sustainable, Profitable Agriculture through Research,” https://eesa.lbl.gov/projects/ar1k-sustainable-profitable-agriculture-research/, accessed October 11, 2019.
13. “ESnet,” http://es.net/, accessed October 14, 2019.
14. “Smart Cities: The Future of Urban Development,” *Forbes*, May 19, 2019, https://www.forbes.com/sites/jamesellsmoor/2019/05/19/smart-cities-the-future-of-urban-development/, accessed October 14, 2019.
15. “Waymo,” https://waymo.com/, accessed October 14, 2019.
16. “Tesla,” https://www.tesla.com/, accessed October 14, 2019.
17. “Earn Money by Driving or Get a Ride Now,” https://www.uber.com/, accessed October 14, 2019.
18. “Hydraulic fracturing Sandia’s role in shale gas production technologies,” http://energy.sandia.gov/wp-content/gallery/uploads/FINAL-HydraulicFracturing-Final-wSAND1.pdf, accessed October 14, 2019.
19. “Hydraulic Fracturing: A Public-Private R&D Success Story,” https://clearpath.org/energy-101/hydraulic-fracturing-a-public-private-rd-success-story/, accessed October 14, 2019.

This page intentionally blank.

DISCLAIMER

This work was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or any third party's use or the results of such use of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof or its contractors or subcontractors. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof, its contractors or subcontractors.