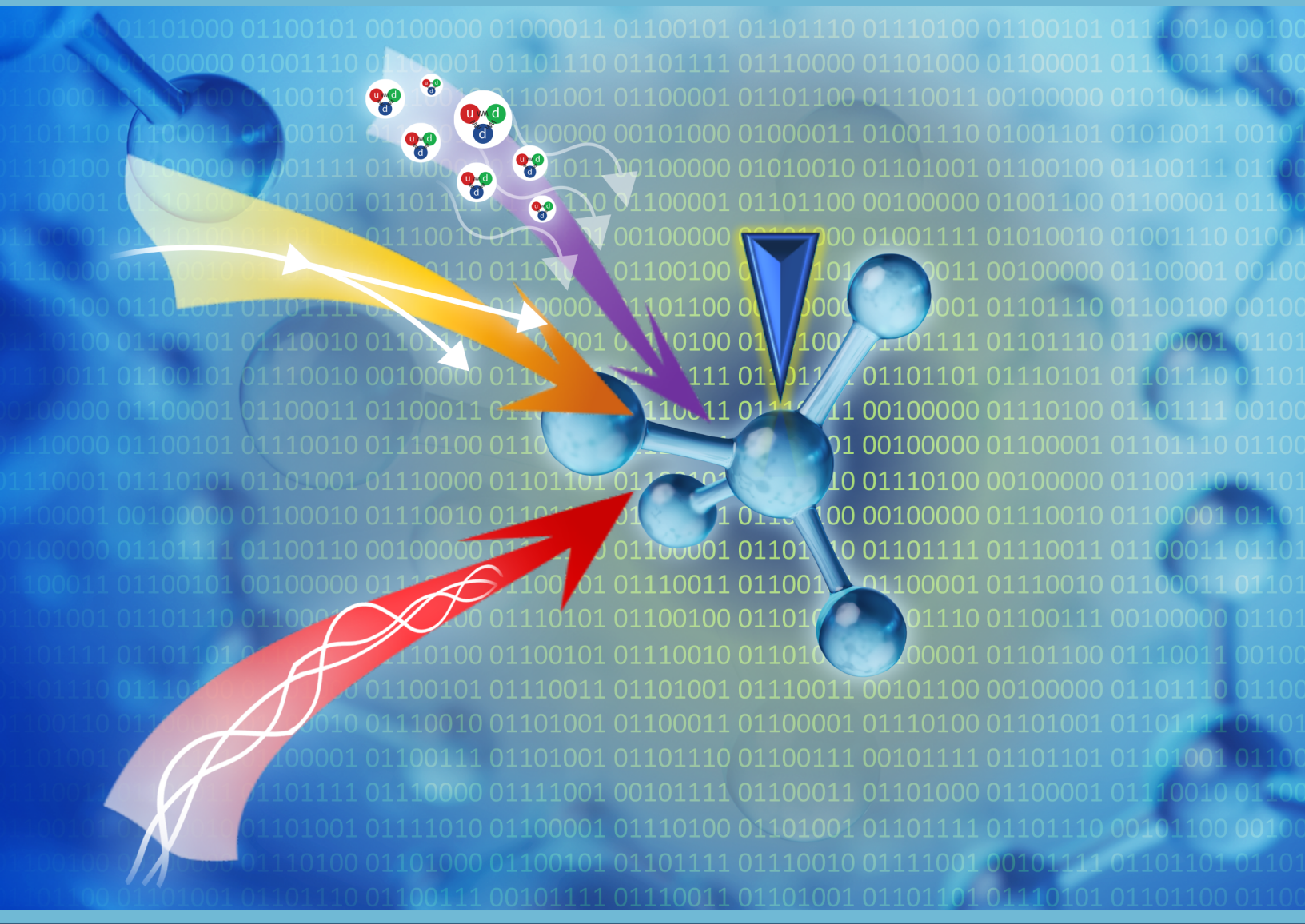
Roundtable on

Producing and Managing Large Scientific Data with Artificial Intelligence and Machine Learning



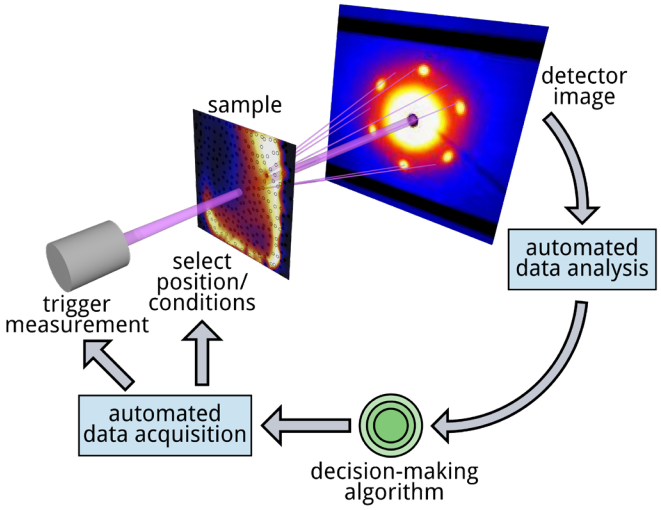
Accelerating experimental and computational discovery through Artificial Intelligence and Machine Learning

Producing and Managing Large Scientific Data with Artificial Intelligence and Machine Learning—Enabling transformative advances at BES scientific user facilities

The US Department of Energy's (DOE's) twelve Basic Energy Sciences (BES) scientific user facilities provide access to the world's most advanced research instruments, annually serving over 16,000 users with impact reported in nearly 7,000 publications and producing unprecedented quantities of scientific data. While impressive, reaching the full potential of these rapidly growing facilities will require new innovations to solve a variety of technical challenges in data acquisition, control, modeling, and analysis. Artificial intelligence and machine learning (AI/ML) have opened corresponding new avenues in optimization, efficient surrogate models, data analytics, and inverse problems. These intriguing capabilities suggest that AI/ML can greatly accelerate the quest to probe and understand fundamental phenomena across a vast range of lengths and timescales, potentially leading to transformative advances across scientific disciplines.

Both industry and science already use AI/ML approaches for data analysis. User facilities, however, crucially require AI/ML tools throughout the lifetime of an experiment: not just for data analysis, but also for data creation, acquisition, and storage. In the next 10 years, AI/ML are expected to go beyond traditional data analysis to aid the design and control of complex facilities, enable real-time capabilities to acquire and analyze large data volumes, automatically steer data collection for in-the-loop experiments, and support experimentalists' use of exascale computing. These advances will in turn open new avenues of scientific research in energy sciences and beyond. For example, we need to transition from relatively simple performance and properties measurements of materials and molecules to complex intertwined functionalities in batteries, information technology, biological systems, and quantum-based devices and sensors, which render classical serendipitous materials discovery and sequential optimization paradigms impractical. We envision a future of AI/ML-enabled facilities that maximize the DOE's scientific impact.

The Office of Basic Energy Sciences held a roundtable in October 2019 to identify coordinated, long-term AI/ML research challenges that will drive major advances in neutron, photon, and nano-based sciences. Four Priority Research Opportunities were identified for the use of AI/ML to greatly enhance the impact of the BES scientific user facilities. The full workshop report will be available at https://science.osti.gov/bes/Community-Resources/Reports.



Autonomous control of experimental systems promises to open the study of problems previously considered impossible. Automating the entire experimental workflow—instrument setup and tuning, sample selection and synthesis, measurement, data analysis and model-driven data interpretation, and follow-up experimental decision-making—will bring about revolutionary efficiencies and research outcomes.

Priority Research Opportunities

- **Efficiently extract critical and strategic information from large, complex datasets**

  Key question: *How do we extract robust and meaningful information from the increasingly vast and complex data now being produced at BES' scientific user facilities?*

  Advances in the tools and techniques available at BES' x-ray, neutron, and nanoscale user facilities allow capture of increasingly larger datasets, often taken in a variety of experimental modalities. Paradoxically, the explosion of data can make it harder to arrive at desired scientific insights because of the monumental level of effort needed to process and analyze the data. AI/ML techniques have the potential to significantly reduce that effort while allowing rapid, real-time information extraction of properties from noisy, imperfect measurements. Additionally, AI/ML can help unmask the complexity hidden in problems in high-dimensional spaces (e.g., multimodal measurements, many experimental variables) by finding connections elusive to human observation.

- **Address the challenges of autonomous control of scientific systems**

  Key question: *How do we address challenges inherent in real-time operation of large, complex scientific user facilities?*

  Realizing the full potential of current and next generation of measurement capabilities will require advanced methods to develop and maintain optimal performance as well as automated experimental approaches to guide scientific discovery. AI/ML-based methods are needed to efficiently search large, complex parameter spaces in real time and to predict the health and failure of instruments that operate at high-power sources and the experiments that are run on those instruments. Such capabilities could dramatically reduce facility tuning time and downtime, improve facility performance, and maximize the productivity of the BES scientific user facilities.

- **Enable offline design and optimization of facilities and experiments**

  Key question: *How do we enable virtual laboratories—offline design and optimization of facility operation and experiments—to achieve new scientific goals?*

  Physically accurate, virtual laboratory environments of experimental facilities (i.e., a lab in the computational cloud) could guide in silico experiments from conception to synthesis and measurements. Digital twins that faithfully mimic facilities, including shared workflows and continuous updates from real experiments, can enable the design of new facility capabilities and execution of optimal experimental strategies to drive physical knowledge acquisition for user facilities. These digital twins could also facilitate development of AI/ML methods for the other Priority Research Opportunities.

- **Use shared scientific data for machine learning–driven discovery**

  Key question: *How can we catalyze scientific discovery by leveraging the wealth of diverse and complementary data recorded across the BES scientific user facilities?*

  Radical improvement in data sharing, curation, and analysis is needed to catalyze scientific discovery across all facilities. Through the application of new AI/ML platforms to integrate diverse scientific data resources, extensive new datasets could be created from heterogeneous experimental and simulated data, leading to new opportunities for scientific discovery. Coordinated development of workflows on a shared facility–based data repository could catalyze development of data standards, formats, and priorities. These curated datasets could, in turn, serve as training sets for developing new AI/ML methods.

Summary

AI/ML methods for data analysis, control, and modeling hold promise for greatly accelerating experimental and computational discovery. Pursuing the Priority Research Opportunities outlined will enable the vision that, in the next 10 years, AI/ML will be an integral part of the discovery and design toolbox, just as experimental, theoretical, and computational tools are today. The BES scientific user facilities can work in synergy with experts across the DOE complex to realize these opportunities and attain the vision of broadly incorporating AI/ML methods in facility operations and scientific experiments. These advances will result in new insights that will drive innovation and enable exploration of scientific space currently unimaginable.

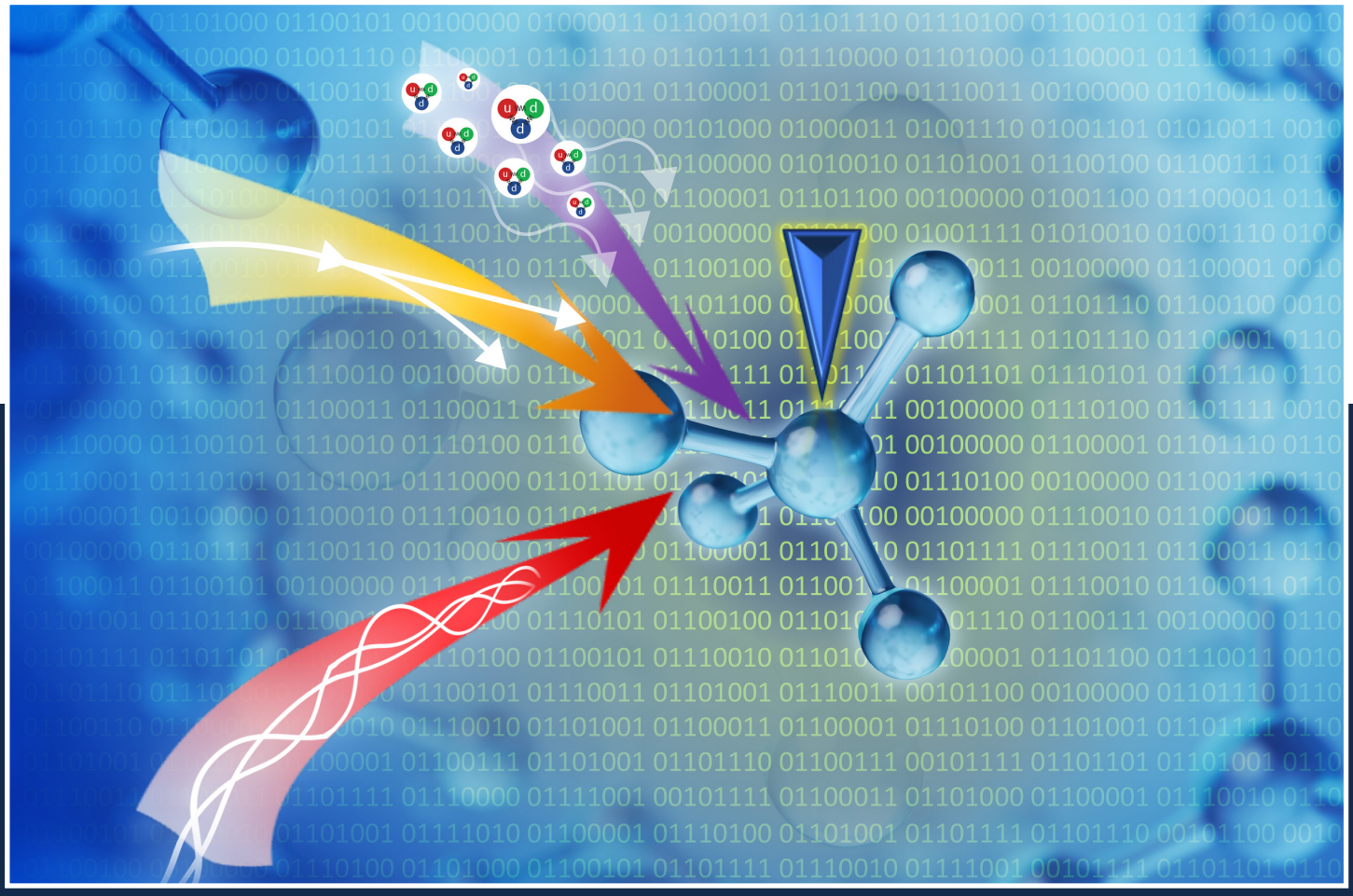


Image courtesy of Oak Ridge National Laboratory.

U.S. DEPARTMENT OF ENERGY | Office of Science