# Big Data Challenges and Advanced Computing Solutions

A Hearing of the

**COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY**
**UNITED STATES HOUSE OF REPRESENTATIVES**

Testimony of

**DR. KATHY YELICK, ASSOCIATE LABORATORY DIRECTOR FOR COMPUTING SCIENCES**
LAWRENCE BERKELEY NATIONAL LABORATORY

July 12, 2018
2318 Rayburn House Office Building
Washington, DC

## INTRODUCTION

Chairman Weber, Chairwoman Comstock, Ranking Members Veasey and Lipinski, and distinguished Members of the Committee, thank you for holding this hearing and for the Committee's support for science. The opportunities presented by "Big Data" are advancing science and innovation in novel and exciting ways. Machine learning is an important part of this story and I commend the committee for exploring how new capabilities in high performance computing and computational science will open doors to new knowledge.

My name is Kathy Yelick and I am the Associate Laboratory Director for Computing Sciences at Lawrence Berkeley National Laboratory, a DOE Office of Science laboratory managed by the University of California. I am also a Professor of Electrical Engineering and Computer Sciences at the University of California, Berkeley. It is my honor and my pleasure to participate in this hearing and to aid the Committee's examination of the opportunities and challenges related to data analytics and machine learning within the Department of Energy. Thank you for inviting me to testify.

Berkeley Lab is a multipurpose lab with world leading capabilities across materials research, biosciences, physics, chemical sciences, energy technologies, earth and environmental sciences, high performance computing, advanced networking and more. Home to five national scientific user facilities, Berkeley Lab serves over 10,000 researchers from all 50 states and beyond. Thirteen Nobel prizes are associated with Berkeley Lab, as are fifteen National Medal of Science recipients. Seventy Berkeley Lab scientists are members of the National Academy of Sciences, one of the highest honors for a scientist in the United States, eighteen of our engineers have been elected to the National Academy of Engineering, and three of our scientists have been elected into the National Academy of Medicine. In addition, Berkeley Lab has trained thousands of university science and engineering students who are advancing technological innovations across the nation and around the world.

In my testimony today I plan to do four things:

**First,** describe some of the large scale data challenges in the DOE Office of Science, drawing examples from Berkeley Lab and other national laboratories' national user facilities and team science projects.

**Second**, talk about the emerging role of machine learning - and specifically deep learning - methods, which have revolutionized the field of artificial intelligence (AI) and may similarly impact scientific discovery.

**Third**, discuss some of the unique opportunities for machine learning in science, leveraging DOE's national role as a leader in high performance computing, applied mathematics, user facilities, and interdisciplinary team science.

And, **fourth,** describe a vision for the national laboratories that includes foundational research in data science along with an interconnected network of experimental and computational facilities to address some of the most challenging data analytics problems in science.

**Part 1: Data challenges in science**

The Department of Energy has a unique role in science as the largest funder of physical sciences research in the nation and with the responsibility for managing and operating many of the largest scientific user facilities. At Berkeley Lab alone, as mentioned previously, there are over 10,000 users of our scientific user facilities, which include the Advanced Light Source, the Joint Genome Institute, the Molecular Foundry, National Energy Research Scientific Computing Center (NERSC), and the Energy Sciences Network (ESnet). In addition, the Lab is a partner in many national and international collaborations, such as the ATLAS, Alice, and CMS projects at the Large Hadron Collider (LHC) in Switzerland, the Dark Energy Spectroscopic Instrument (DESI) near Tucson, Arizona, and the LZ dark matter experiment in South Dakota.

Big data challenges are often characterized by the *4 Vs*: volume (the total size), velocity (the speed at which is it being produced), variability (the diversity of data types) and veracity (noise, errors, and other quality issues). Scientific data has all of these, and DOE's user facilities are a big source of the challenges and the opportunities to use large data sets for new discoveries, because of increasing data rates, reduced costs of collecting data, and total data volumes.

The cost of sequencing the human genome is now around $1,000 down from $10,000,000 just a decade ago, and the National Institutes of Health (NIH) database on genomic data (the Sequence Read Archive, or SRA) now holds over 8 petabytes ($10^{15}$ bytes) of genomic data, a 3000x increase in ten years. At DOE's Joint Genome Institute, a newer database of viral genomes has grown nearly 100x in just two years.

In cosmology and particle physics, the velocities and volumes have also grown, with the upcoming Large Synoptic Survey Telescope (LSST) producing about 20 terabytes ($10^{12}$ bytes) every night and a resulting community data set over its lifetime of about 60 petabytes. The LHC will collect roughly 50 petabytes of data in 2018, even after eliminating 99% of the data produced inside the experiment, and that 50 petabytes will grow to roughly 500 petabytes by 2024. The LHC data from past experiments is copied to data centers around the world with 900 petabytes currently in disk and tape storage.

The volume and velocity of scientific data is growing because the instruments are improving -- we can see things at a microscopic and atomic scale, measure vibrations imperceptible to the human eye, and take high resolution images of objects in the universe that are millions of light years away. The national labs are key to developing many of the instruments used for major science experiments. For example, Berkeley Lab has a long history of developing the detectors used in electron and x-ray microscopy, improving spatial resolution 100-fold and temporal resolution 1,000-fold, to reveal atomic structure without the need for crystallization. This technology was revolutionary in chemistry, material science, and biology, and its use in Cryo Electron Microscopy instruments was cited in the 2017 Nobel Prize in Chemistry, as well as being commercialized for advanced medical and scientific imaging.

Data veracity and variety are also challenges in nearly every discipline in science, with scientists eager to extract vanishingly small signals from large messy data sets and combine different modalities to improve insights. One of the most exciting fields for the application of big data science and advanced computing is biology - the increasing sizes of data sets, the inherent noise, and the complexity make it a prime area of research to leverage these emerging tools and capabilities.

Microbes are a data challenge - they are the most abundant and diverse life form on Earth. They exist in vast complex microbial communities called microbiomes and interact with and significantly impact all of the world's natural systems, including human, plant, animal, energy, and environmental, at all scales, from the infinitesimal to the grand. In one handful of soil there may be as many microbes as there are stars in our galaxy. Discovering how these complex communities of millions and billions of microbes interact and impact natural systems will create new knowledge and advance solutions to the world's most intractable problems - unlocking and harnessing the mysteries of microbiomes will advance environmental remediation, propel new agricultural technologies and processes, and speed biologically-based energy solutions to market. This research will grow the United States' bioeconomy and drive tremendous economic activity. Maintaining U.S. leadership in microbiome research is an economic and national security strategic imperative - but, it's a hard nut to crack, in large part due to the data challenges.

For example, a particular microbial species and those its genetic data often occur only in samples with hundreds of different species and thousands of strains mixed together. To further complicate analysis, today's standard sequencing technologies produce error-laden fragments of DNA that need to be assembled together to find genes. Putting it all together in a scientifically useful way is analogous to completing hundreds of different jigsaw puzzles with all the pieces mixed together, some of the pieces broken, and no reference pictures for any of the final images. Of course, the function of a microbiome is more than the sum of the parts, with multiple species interacting to impact the environment in which they live, whether it's within the human body or in the environment. To understand and eventually control microbial behavior from the genomic level, one also must combine genomic data with a variety of data from imaging, chemical sensors, and other scientific instruments - making an already complex task more difficult.

High performance computing and novel computational methods give scientists the tools needed to decipher these microbial puzzles and to assemble, shape and coax the data into useable information, removing errors, finding genes, and discovering relationships between species and across different microbiome samples. My own research includes leading the microbiome application project in the Exascale Computing Project (ECP), where the overarching goal is to find new information about the microbiome using more powerful algorithms and computers. It may reveal changes in the microbiome due to diet, weather, chemicals or other environmental factors, and ways of using a microbial community to produce a healthier environment for humans as well for food crops. More broadly, scientists have recognized the need for interdisciplinary research efforts in this area and for a National Microbiome Data Collaborative, an open, standardized, and shared data infrastructure, that could help foster integrated analyses and synthesis across diverse microbial datasets.

As another example of an ECP application, scientists and engineers at Berkeley Lab and Lawrence Livermore National Laboratory (LLNL) recently performed a simulation of a large-magnitude earthquake in the San Francisco Bay Area, and how it would affect different locations and buildings, with the goal of understanding impacts on critical infrastructure such as schools, hospitals, and the power grid. This was done at an unprecedented scale and resolution using Berkeley Lab's NERSC supercomputers. Even larger simulations will be done on future exascale systems. Already, these simulations have shown that the same building located less than 3 miles apart may have different risks and therefore require different building hardening.

To make the results even more specific to a given location, the team is looking at using measured seismic data obtained from regionally deployed sensors during frequently occurring small earthquakes to help improve fine-scale geologic models. By computationally "inverting" measured data, enhanced understanding of the subsurface geology can be obtained to improve the computational models for ground motion simulations. Merging high performance simulations with measured data will yield even more precise information about shaking at every location throughout the region. While this massive aggregation is no small undertaking, it will lead to improved public safety. These types of approaches are only becoming feasible because of the major advancements being made in high performance simulation combined with big data exploitation.

Looking towards the future, even more dense data will become available as seismic sensors proliferate. For example, recent technology advancements provide an opportunity to deploy seismic sensors across the electric grid onboard smartmeters, which will provide unprecedented data. We need to be prepared to exploit this emerging big data for transforming hazard and risk assessments.

**Part 2: Data Analytics, Machine Learning, Deep Learning, and Artificial Intelligence**

***Machine learning*** is an increasingly popular strategy for analyzing scientific data and offers opportunities to better leverage and benefit from the explosion in data volume. It's also a term that is used very broadly to refer to methods that learn from data, or to make inferences based on a model learned from some data. The most well-known example is identifying images, such as cats, on the internet. A machine learning algorithm is fed a large set of, say, 10 million images of which some are labeled as having a picture of a cat. The algorithm uses those examples to build a model of which images contain cats, i.e., the probability that a given image contains a cat. For example, an image with two diagonal lines that meet to form something like a cat's ear will have a higher probability of containing a cat. This is known as *supervised learning*, because one starts with images labeled as cats or not, whereas *unsupervised learning* might have a set of unlabeled images and can determine which ones have similar objects in them, but does not determine what the objects are. In science we are not looking for cats, but we can use machine learning to find features such as exploding stars, cellular structures, or subatomic particles.

There are several different kinds of machine learning algorithms, but many of the most notable breakthroughs in recent years have come from a powerful class of *deep learning* algorithms, and

specifically *deep neural networks*, which are used in this example of finding cats or other images in internet searches. The algorithm works in a set of layers, where one layer may find the edges between different objects in a picture, another layer will find shapes from the edges, and higher level layers will find recognizable objects such as eyes, ears, and tails. The number of layers will vary depending on the application problem, and it is one of the things that a data scientist may have to experiment with, but typically the depth is a few layers to a few dozen.

Deep learning has led to a number of surprising results in the field of Artificial Intelligence (or AI), which has the goal of developing computers with human-like capabilities, including computer vision, speech recognition, robotics, and playing games of strategy. Deep learning is used in Siri to recognize speech and interpret commands, and it is used in self-driving cars to identify road signs, hazards, and obstacles. It was also used by Alibaba, a Chinese company, to outperform students on a standardized reading comprehension exam, similar to what is used in college admissions, and by Google's AlphaGo in 2016 to beat the world ranking world champion (Lee Sedol) in the game of Go, a strategy board game that is significantly harder than chess. These deep learning methods are so strongly linked with AI that the terms *AI* and *deep learning* are sometimes used synonymously.

These ideas have been around for a long time -- the neural net ideas go back to the 1940s, and the key algorithmic idea (*backpropagation*) was developed in the 1980s. So, why have these algorithms suddenly become successful? Roughly speaking, machine learning requires three things: large amounts of data, fast computers, and good algorithms. The growth in data has been fueled by the ubiquity of cameras, recording devices, and various sensors, facilitated by the ease of sharing data over the Internet, while computing performance has grown by a factor of one million since the early 80s. As described earlier, there has been a similar explosion in data in science coming from instruments that provide more detail and higher data rates, and from increasingly complex simulations enabled by faster computers. With DOE's abundant use of simulation, faster computers are both part of the challenge and part of the solution.

DOE's unique resources in high end computing have also proven to be well suited to machine learning, and deep learning maps well onto the Graphics Processing Units (GPU) in the pre-exascale systems recently deployed in the Summit machine at the Oak Ridge Leadership Computing Facility (OLCF) and Sierra at LLNL. One of the key computational kernels in deep learning is multiplying two matrices, which also is the dominant kernel in the Linpack benchmark used for the TOP500 list, where Summit and Sierra are in the #1 and #3 spots respectively. Not surprisingly, some of the early science projects on these computers are focused on machine learning, and are using the high speed networks, large amounts of memory, and unprecedented computing capability to solve problems that are intractable on conventional computers.

Each year the top performing scientific application team is awarded the Gordon Bell prize, an award that reflects science at scale, as opposed to a fixed benchmark. Finalists for the 2018 prize

include a deep learning computation at over 200 petaops[1]/sec computation on Summit, which was a partnership between NERSC, OLCF, NVIDIA, and Google, that was used to analyze data from cosmology and extreme weather events. A second finalist is a project lead by Oak Ridge National Laboratory with researchers from the University of Missouri in St. Louis, which used an entirely different algorithm to learn relationships between genetic mutations across an enormous set of genomes, with potential applications in biomanufacturing and human health. This algorithm can also be mapped to matrix-multiply like operations. It runs at a impressive 1.88 exaop/second! These are the fastest deep learning and other machine learning computations to date.

**Part 3: The use of machine learning in science**

**What does this mean for science?** The image analysis example is directly analogous to science, because images arise in many scientific disciplines, from electron microscopes in biology, to x-rays from light sources used in material science, to telescopes used in cosmology. Saul Perlmutter, a Nobel Laureate from Berkeley Lab, used image analysis to discover the accelerating expansion of the Universe through observations of distant supernovae - exploding stars - as a kind of standard reference point. His team used a specific kind of supernova, Type 1A, which occurs when a white dwarf star explodes; these are fairly rare, with about one per century within our Milky Way Galaxy. Using high powered telescopes and collecting images over many months, they would look for the appearance of new stars in remote galaxies that suddenly appeared on an image (called a "transient") and then use other telescopes, like the Hubble Space Telescope and the Keck Telescopes, to classify the transient as a variable star, a quasar, or supernova.

Thirty years ago, a few tens of images were produced each night and were analyzed manually by scientific experts. By 2007, some automatic processing was done to find transients, and Berkeley Lab was already working on using machine learning algorithms to classify supernovae. Today, tens of millions of images are produced from experiments like the Dark Energy Survey, the Zwicky Transient Facility, and soon the Large Synoptic Survey Telescope, which produce thousands of new transient discoveries each night. Machine learning algorithms run automatically on supercomputers at NERSC and the National Science Foundation's National Center for Supercomputing Applications center in Illinois, scouring these images each night for new transients. These machine learning algorithms make scientists much more productive, reducing by more than 10,000x the number of images they look at manually. Today, the focus is on using deep learning to not only find these transients, but to classify them so that follow-up resources are only spent on those objects the scientists want to study.

---

[1] Petaop/second and exaop/second are, respectively, $10^{15}$ and $10^{18}$ operations per second. For machine learning applications, the computations can often be performed using less powerful operations than normal (double precision) floating point operations (*flops*) used in High Performance Computing (HPC). These machine learning "ops" are about one quarter as powerful those typical HPC flops.

Machine learning can often find patterns in noisy data when other approaches fail, so scientific applications are not limited to cosmology or even to images. We recently surveyed researchers at Berkeley Lab and found over 100 projects, many in partnership with other labs and universities, that are using some form of machine learning. For example, researchers from Fermilab, Caltech, and Berkeley Lab are exploring the use of deep learning to identify and track particles in experiments at the Large Hadron Collider, working to replace current algorithms that are not easily implemented on high performance computers using traditional approaches, and are projected to consume enormous amounts of computing time after the LHC upgrade. Another group has developed machine learning strategies that aim to increase crop yields and improve the sustainability of agriculture while reducing economic risks for farmers and landowners as part of the AR1K collaboration between Berkeley Lab, the University of Arkansas, and Glennoe Farms. The farm is an experimental platform instrumented with sensors, drone-based imaging, and frequent data collection, and machine learning is the tool that will tie all the data together. The Lab's Center for Advanced Mathematics for Energy Research Applications (CAMERA) has developed a new deep learning algorithm to analyze light source images more quickly and more accurately than previous approaches, and the Materials Project is using machine learning to remove the guesswork from materials discovery and design, driving the development of advanced materials.

At DOE's Joint BioEnergy Institute (JBEI), scientists are using machine learning to improve biosensor design and accelerate synthetic biology to produce biofuels; the technique can predict the amount of biofuel produced by newly engineered bacterial cells based on data from previous experiments and has other applications, such as developing drugs that fight antibiotic-resistant infections and crops that withstand drought. And at the Joint Genome Institute, machine learning is used to answer fundamental questions about biology, such as the relationships between all genes that naturally occur in the environment.

DOE researchers are also using machine learning to improve the operation of its facilities and make scientists more productive. An enormous challenge in large data is getting labels or metadata information associated with data from each scientific experiment - making it more accessible and useful to scientists. Researchers at Berkeley Lab have developed techniques to automatically label data, starting with the enormous stream of data on advanced materials for batteries and other applications, coming from the National Center for Electron Microscopy (NCEM), home to the world's most powerful electron microscope. In another example, ESnet, DOE's advanced high speed scientific network, is using a variety of machine learning techniques to predict the amount of data being transferred between endpoints in order to adapt network traffic dynamically, detect problems in the infrastructures, and find anomalous high-volume data transfers, which could indicate either a faulty device or a cyber attack.

Similar ideas are used for other real-time flows of time-series data, such as looking for abnormal behavior in the power grid, in rooftop solar panel systems, or even financial market data. Data from cell phones and embedded sensors are being used to build large-scale models of regional transportation systems, which can be used for long term planning of transportation infrastructure, energy planning, and emergency response.

Machine learning expertise at the DOE labs can also be leveraged for other national priorities, as in partnerships with the NIH and the Department of Veterans Affairs (VA). The data in this case includes genomes, images, results of medical tests, and electronic medical records. It is being used to address medical challenges such as traumatic brain injury, cancer, mental health, and the opioid crisis.

Berkeley Lab researchers have developed machine learning approaches to analyze and visualize brain image data collected from multiple devices, to automatically identify cancer cells in image data, fibers in textile images, and more. DOE researchers bring expertise on high performance computing, data analytics, modeling and simulation, as well as a culture of team-based science, where the team of cross-disciplinary scientists and engineers work on end-to-end solutions for grand challenge problems.

Machine learning does not replace the need for the more traditional use of HPC simulations, but instead offers a complementary set of techniques. Roughly speaking, simulation is used when a set of equations, i.e. a model, is known in advance, while machine learning may be used to infer a model based on data from observations. Machine learning is often combined with simulation to fill in parts of simulation where no known equations exist but where data is available. This approach is being used for simulating turbulent fluids by researchers at Sandia.

Machine learning is also used to accelerate large ensembles of simulations, where the machine learning can quickly approximate them to determine which ones are most important in searching for a particular outcome. Finally, machine learning can be performed on the data produced by simulations, such as in research at Berkeley Lab searching for extreme weather events. As stated in a recent report by the DOE Advanced Scientific Computing Research (ASCR) community on scientific machine learning, “In all cases, it is clear that ML will not replace decades of research in principled physics-based approaches. Rather, it can bring a toolbox of methods to enrich, improve, and accelerate current modeling approaches.”

**Part 4: The need for foundational research and interconnected facilities to advance data-driven scientific discovery**

As indicated from the examples above and many more across the national lab complex, scientists are actively pursuing the use of machine learning and advanced analytics in nearly every basic and applied scientific domain. Enthusiasm is appropriate based on existing results from AI and from the success of many commercial applications. However, enthusiasm should be tempered with some understanding of the challenges facing scientific applications. ASCR’s recent workshop on scientific machine learning elucidated many of these challenges and the need for additional research on the mathematical foundations of machine learning, including the following topics:

1) Leveraging scientific domain knowledge: Many machine learning techniques have been developed primarily for images, speech, and textual data. Speech and textual data may have use in analyzing scientific publications, notebooks, and presentations, but would not be the

primary focus. And, while images are common in science, their formats and content will be quite different than in more common internet image searches. Much of scientific data from simulations is three dimensional and may exhibit symmetries not recognized by current algorithms, e.g., in molecular structures. Finally, scientific data is often incomplete and acquiring large sets of labeled data for supervised learning may not be practical.

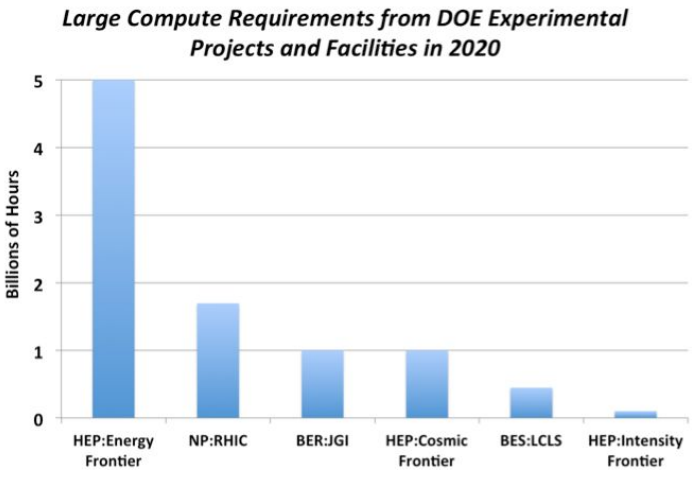
2) Interpretable machine learning: The models derived from machine learning methods, and especially deep learning, are not a recognizable set of equations but instead may involve many thousands of parameters without an explanation of what each one means. For placing advertisements or classifying images, this may be acceptable, as long as they make good choices, but scientists will demand more confidence, better error bars, and in some cases direct explanations for use in developing theories. Simple correlations are not sufficient.

3) Robust and efficient: There should be rigorous numerical estimates for the quality of results and well-defined criteria on when the inferences may be used. When applying machine learning to infer properties of data, one distinguishes between that data given to the algorithm for "training" and the data on which it will be used. For example, images labeled as containing a supernova may be given as training, and once trained, it can be used to automatically label images. Methods should be insensitive to the details of how training data is selected and should perform well as long as the training set is in some sense reasonable. In AI research, there are concerns about bias that derives from the selection of training data -- only posting advertisements for a CEO position to white men, for example. In science, bias may come from artifacts of a particular instrument or measurement technique.

The ASCR workshop report also recognized the close interaction of simulation and machine learning, as well as opportunities to control large scientific campaigns, whether a large set simulations or experiments, to choose the best cases to run.

While the workshop was targeting mathematical research in machine learning, there are also computer science problems related to programmability, performance, parallelization, and scale. There is currently a strong preference in the deep learning community for GPU architectures, but even more specialized architectures may prove beneficial, which could create a divergence between the computing platforms for simulation and learning. Other machine learning algorithms may place higher demands on the memory system and network, and as the foundational work evolves, deep learning methods may use sparsity, e.g., to improve running time or interpretability, which will also shift hardware requirements.

DOE will need to address the burgeoning data analytics needs from major experiments and embedded sensors, whether those use deep learning, other machine learning methods, or other analytics techniques. The figure on the right shows the



estimated computing requirements of some of the major DOE experiments, normalized to NERSC computing hours. For comparison, NERSC currently delivers about seven billion hours to its users. Some of the data analysis will require real-time processing, automated job scheduling, and other policy changes, in addition to sufficient computing to meet this growing scientific demand.

**Conclusions**

Data-driven scientific discovery is poised to deliver breakthroughs across many disciplines, and as stewards of many national user facilities, DOE should have a leadership role. Driven by innovations in instrumentation and computing, and a desire to investigate increasingly complex scientific questions, the data challenges will continue to grow. In addition to analytics problems, there are many technical challenges in data curation, sharing, metadata labeling, and search, to give scientists tools for research that are analogous to those that have revolutionized shopping, entertainment, and business.

Machine learning is a promising approach for analytics in science, complementing but not replacing modeling and simulation. In spite of the extensive work already going on at the DOE labs, machine learning and associated mathematical foundations of machine learning are not as well developed as in simulation science.

The goal in scientific discovery is more focused than so-called *general AI*, but also goes beyond emulating human capabilities. The scientific instruments described earlier are the eyes, ears and hands of science, and the goal is not to replicate human behavior but augment it with superhuman measurement, control and analysis capabilities, empowering scientists to ask and answer more complex questions. For this reason the alternate phrase, *Intelligence Augmentation (IA),* is probably more appropriate.

The Exascale Computing Initiative is addressing one of the three requirement to make machine learning successful in DOE: availability of extreme computing capabilities. The Exascale Computing Project is addressing some of the underlying computational challenges of data analytics, with applications in cancer research, microbiome analysis, and light source imaging, all involving some form of machine learning along with other simulation and analytics methods. The project also has co-design centers in graph analytics and machine learning, which are linked to a number of the 24 applications. Along with the procurement strategies at the computing facilities, ECP will ensure that future exascale system architectures are well suited to this workload. But the foundational research in machine learning and broader facility issues still needs to be addressed.

While raw data is growing, DOE will need a strategy and infrastructure to enable sharing, search, curation and management, and to ensure the facilities are coupled in a way that large experiments can take advantage of the high end computing facilities for the largest data challenges

In the excitement over machine learning methods and data produced by new instruments and embedded sensors, one should not lose sight of the need for stronger foundational work. DOE

has taken this seriously in the field of modeling and simulation, developing mathematical models and algorithms with proven performance and quality metrics, along with quantifiable measures of uncertainty and errors. The scientific peer review process will drive machine learning to be similarly rigorous, in a way that the commercial applications do not. DOE's interdisciplinary approach, which will require additional expertise in statistics and mathematical optimization, in addition to current strengths in applied mathematics and computer science, should lead to high quality methods that solve real problems and lead to new methods and insights that will benefit other applications of machine learning.

Katherine (Kathy) Yelick is a Professor of Electrical Engineering and Computer Sciences at UC Berkeley and the Associate Laboratory Director (ALD) for Computing Sciences at Lawrence Berkeley National Laboratory. Her research is in high performance computing, programming languages, compilers, parallel algorithms, and automatic performance tuning. She currently leads the Berkeley UPC project and co-lead the Berkeley Benchmarking and Optimization (Bebop) group. As ALD for Computing Sciences at LBNL, she oversees the National Energy Research Scientific Computing Center (NERSC), the Energy Sciences Network (ESnet) and the Computational Research Division (CRD), which covers applied math, computer science, data science and computational science.