Testimony of Dr. Dimitri Kusnezov

Chief Scientist

Department of Energy/National Nuclear Security Administration

Before the

Subcommittee on Research and Technology & Subcommittee on Energy

Committee on Science, Space and Technology

U.S. House of Representatives

May 22, 2018

Thank you Chairwoman Comstock and Chairman Weber, Ranking Members Lipinski and Veasey, and distinguished Members of the Subcommittees. I thank you for taking up this important issue and for the opportunity to address the Members and share what the Department of Energy (DOE), in collaboration with the Department of Veterans Affairs (VA), is trying to do at the intersection of next-generation supercomputing architectures, U.S. innovation and veterans' health.

Introduction

Driven by complex and often urgent national missions, DOE has honed an ability to attack short- and long-term challenges through big team science and technology efforts. Supported through the technical base of our national laboratories as well as academic, private sector and international partners, no challenge has been too big. In the past three years, several important moments created opportunities for innovation in support of national goals. This includes the National Strategic Computing Initiative (NSCI) in 2015, the Cancer Moonshot in 2016, the 21st Century Cures Act, also in 2016, and Secretary Perry's commitment, as well as this Administration's commitment to Veterans' issues. These have forced the rethinking of traditional paradigms and facilitated novel approaches to how we solve complex problems.

Driven by the aforementioned initiatives' demand for Federal collaboration, in early 2016 a cooperative relationship between DOE and the VA was established. In 2017, this evolved into a Big Data Science Initiative (BDSI) which currently encompasses the Million Veterans Program–Computational Health Analytics for Medical Precision to Improve Outcomes Now (MVP-CHAMPION). As is the case for many of the DOE's missions, next generation computing is making possible things previously thought to be impossible. The willingness to attack complex and significant problems is a hallmark of DOE.

This collaboration addresses both the mission spaces of the VA in treating and improving the lives of veterans, and the expertise and facilities residing at the DOE labs, and through the Economy Act, VA can work with DOE. In the FY19 VA Budget Request, $27 million identified to support this cooperative relationship.

DOE roles in the Life Sciences

It is worth taking a brief moment to recognize DOE's long history of working in the biosciences. Rooted in our history, from the dawn of the nuclear age, was the development of extensive radiation biology programs at the DOE national laboratories. Today we have to manage risk of exposure in everything from the remediation of legacy facilities and waste to planning new construction. Expertise in the biosciences remains core to our responsibilities, however today that is focused on biology to tackle problems from energy production, to determining genomic properties, molecular and regulatory mechanisms, and the resulting functional potential of microbes, plants, and biological communities central to DOE missions.

The Human Genome Initiative, conceived by DOE in 1986, is one of the most transformational efforts in the life sciences. While it built on the existing genome programs at the DOE labs, it was the strength of the labs in adjacent fields that changed a research enterprise from being almost exclusively hypothesis-driven to a data-driven and tool-driven one. It was the resident expertise in these adjacent fields that recognized that advancements in robotics, dye-tagging, computing, lasers and so forth, could be drawn together, converged, to create a viable, yet ambitious, path to sequence the human genome. The initiative was transformed into the Human Genome Project (HGP) in 1990, drawing in the National Institutes of Health (NIH) to help drive the program and engage its research base.

What we know today is that this effort created an entire new economic sector. The cumulative costs through 2003 were likely several billion dollars, as estimated by investments of all the partners, for delivering that first sequence. By the turn of the decade, the global economic impact was already evident, and the benefit to the U.S. economy alone was estimated to be $141 for every $1 spent in the project[1]. Today sequencing is routine and done for many reasons from learning about one's genealogy to forensic sciences or precision medicine. The cost has come down by over a factor of one million, with $100 sequencing on the horizon. We may be at a similar watershed moment in precision medicine and next generation supercomputing, and efforts focused on Veterans' health could be equally or more transformative.

High Performance Computing at Turning Points

The Department, and its predecessors, have tallied over 70 years of leadership in supercomputing and solving big, complex problems. As with the life sciences, our

[1] Tripp, S. and M. Grueber, *Economic Impact of the Human Genome Project: How a $3.8 billion investment drove $796 billion in economic impact, created 310,000 jobs and launched the genomic revolution*, Battelle Memorial Institute, May 2011.

expertise is rooted in the Manhattan Project, and the demands of the mission has stressed the very definitions of computer designs and uses to, not only, meet the nation's needs but redefine the scientific landscape. We have been known for computing at scales that have placed the U.S. in a position of supercomputing pre-eminence for decades. These supercomputers have been and continue to be the cornerstone of our scientific stockpile stewardship program that assures the safety and reliability of the U.S. nuclear weapons arsenal. It is a core competency of DOE and increasingly applied to our missions, from our energy sector responsibilities, to nuclear and cyber security, the grid, fracking to more effectively liberate fossil fuels, and generally as a means to better inform complex decisions or support through scientific discovery.

There are foundational changes coming to this sector that DOE is currently managing through our Exascale programs in the National Nuclear Security Administration and DOE's Office of Science. Exascale is an effort to continue along the high-performance computing trajectory long defined through Moore's Law, but now only after significant performance issues are resolved. These include requiring vast improvements in power efficiency, resilience, computer memory and parallelism. We expect these systems in the next several years, building on the pre-exascale systems currently being delivered at Lawrence Livermore, and Oak Ridge National Laboratories. These systems are important because they provide the needed continuity for our missions to deliver on tools long validated to their problems of interest.

At the same time, there has been notable growth that could be termed a "big data" revolution, leading to many advances, applications and a growing economic sector focused on applications of data science. At DOE we have made contributions to this revolution. What is challenging is the growing scale of data and simulation data.

What underlies effective numerical prediction on our high performance computers (HPCs) is how to deal with big data through uncertainty quantification (UQ). This is our approach of ascribing how much we believe in predictions from the computer and ultimately how we make decisions on critical national security issues. Every day we use UQ for our nuclear weapons missions where certainty in predictions is paramount. Analogous to the hurricane model predictions we see annually with the cones of expected landfall impacts, the discipline developed within DOE is far more complex, but increasingly challenged by the remarkable growth in data capabilities. Imagine having one billion high-resolution photos on your smartphone and asking yourself what kinds of knowledge may be contained in these images. This is the challenge of petabytes of data. Imaging another thousand times more. We are generating experimental data routinely at these scales, and running complex numerical simulations producing equally large amounts of numerical data. Rooted in our scientific method, is the need to state with some measure of precision, the confidence we have in our predictions. Today we have no means to meld the vast amounts of experimental data with the complex simulation data to yield measures of our confidence. Much, if not most, of all of this data is simply not used because it is beyond our ability to fully appreciate or comprehend. We extract what we believe are the most salient features of both experimental and numerical data and use that

comparison as our figures of merit. Our computing paradigm cannot accommodate these needs.

The NSCI launched in July 2015 defined a federal strategy to meet the nation's needs and assigned leadership in exascale to DOE. It identified DOE as a lead agency for high performance computing and defined goals such as adopting "a whole-of-government approach that draws upon the strengths of and seeks cooperation among all executive departments and agencies with significant expertise or equities in HPC while also collaborating with industry and academia." It also required that DOE establish "a viable path forward" for Post-Moore's Law computing. Consequently, over the past several years, we have been investing in Post-Moore's Law technologies, from brain inspired neuromorphic computing, artificial intelligence (AI) and machine learning (ML), to quantum information technologies.

I use AI and ML as the broad umbrella of hardware and software approaches we use today to ask ourselves what we can learn from large and complex data sets without having to specify what we are looking for. It serves as a means to surface often complex or subtle relationships in data, surfacing patterns, and identifying simpler representations of the data.

This last point is especially crucial, and is at the core of why the partnerships in precision medicine and with the VA are so important. The growth of brain inspired computing has been tremendous over the past several years. From Google to Amazon and Walmart, progress is evident almost everywhere. Even the iPhone X has adopted neural inspired technologies into its functionality. This aspect of AI is becoming increasingly prevalent. We began to recognize that AI might be the only means to reconcile the exponential growth in data with our needs to make predictions. Given our longstanding investments in the current computing paradigm in which UQ is a post-hoc addition to the computation, the challenge was finding the means to:

- Force the rethinking of traditional computing paradigms by challenging scientists with qualitatively new classes of prediction and a richness of data;
- Use the qualities of data to change how we think of many of our traditional approaches from computer architectures to UQ to codes.

Precision medicine became the evident accelerant to drive this technology branch of computing, which we would term as the convergence of AI, big data analytics, and our traditional HPC. It has become a means to reconcile the problem of too much complex data, data beyond human abilities to process, and the need to predict.

In December 2016, the 21st Century Cures Act was passed into law. It recognized the important role that DOE should play in precision medicine. Specifically, in TITLE II—DISCOVERY Subtitle B—Advancing Precision Medicine, SEC. 2011. PRECISION MEDICINE INITIATIVE amended Part H of title IV of the Public Health Service Act 19 (42 U.S.C. 289 et seq.) to now include a role for DOE to identify and address the advanced supercomputing and other advanced technology needs for precision

medicine. The convergence of technologies and health sciences will define where precision medicine goes and provides an opportunity for a transformational shift in bioscience. To take this transformational leap, precision medicine is going to drive the technological convergence of traditional supercomputing, smart computing (advanced machine learning and AI), and secure computing, an activity for which DOE is uniquely qualified. This supports the VA FY19 Budget Request and it's intent to engage the relevant services and expertise at the DOE labs.

DOE, VA and MVP-CHAMPION

During 2016, I co-led the Data and Technology Track for the Cancer Moonshot. One of the enduring challenges was that while there is big data out there, it is nearly impossible to access any of it. The partnership with the VA was likely the largest unleashing of medical data to come out of this. While the effort goes well beyond cancer, it originated in the early moments of the Moonshot while the NSCI was fresh and there was an identified incentive to find partnerships to open up data and advance the missions of agencies through HPC.

The partnership between the VA and DOE was engineered to advance the priorities of both agencies. The burgeoning field of data science that includes techniques such as AI has been rapidly transforming our economy at large but is increasingly responsible for advances in biomedicine. These new generation of data science tools grow increasingly more powerful as the breadth, depth, and complexity of the dataset increases. The most striking aspect of this revolution has been the ability of these tools to detect signals that have been stubbornly unapparent to human inspection or otherwise undetected with current data science approaches. One of the challenges of caring for our veterans is the uniqueness of the service-connected conditions derived from serving in the U.S. Armed Forces that are not frequently encountered in civilian medicine. Detecting such types of signals has been a hallmark of AI. The VA has a unique dataset of medical records, whole genomes, and imaging data that is one of the most comprehensive in dimensions of time, scale, and breadth, and in many aspects, this dataset is considered to be the largest and most comprehensive in the world.

The VA-DOE partnership began in early 2016. Following some calls and meetings between DOE, VA and the White House, we pulled together a small team from four DOE labs who were willing to take initial steps to develop this partnership on short notice. Representatives from Lawrence Livermore, Los Alamos, Oak Ridge and Argonne National Labs met with the VA at their Boston, Jamaica Plain site and explored the technical challenges faced by the VA in the ambitious and unparalleled Million Veterans Program. This developed and became MVP-CHAMPION (**M**illion **V**eteran **P**rogram **C**omputational **H**ealth **A**nalytics for **M**edical **P**recision to **I**mprove **O**utcomes **N**ow) in mid-2016. We marked the start with a Statement of Principles co-signed by the Energy Secretary and VA Secretary in June of 2016. Since then, we have added many additional documents to this, from governance charters, the Business Associate Agreements, Institutional Review Board (IRB) agreements, Data Use Agreements, Rules of Behavior and shortly a MOA.

A Big Data Science Initiative (BDSI)

The growth of MVP-CHAMPION from its initial starting point of genomic data coupled to electronic health records into an even more data rich effort is captured in our identification of this as a BDSI. This developed in April of 2017 as we convened to rescope and further define our goals within this initiative. Data transfer activities initiated in November 2016, through a reimbursable IAA for approximately $3.4 million, moving data from the VA to DOE. Today we are envisioning mirror sites, including Argonne and Livermore, where different versions of the data can be configured to tackle distinct classes of problems, including TBI. These sites are significant, since with Oak Ridge, they are pre-exascale and exascale sites where next-generation supercomputer architectures are resident and can be engaged. We are intending to gather VA data, VA provided data, and other sources of data. These include sources such as:

- Million Veterans Program (MVP), currently the world's largest collection of genotypic data;
- VA Corporate Data Warehouse;
- VA disease specific registries;
- National Death Index;
- Center for Medicare and Medicaid Services and other pubic and academic data collections;
- Other VA data types including expanded sets of 'omics' data (e.g., DNA, RNA, proteins, metabolites, microbiome);
- Clinical images (for example radiology, pathology, other medical images);
- Patient-generated data (patient wearable medical devices as an example);
- Social and environmental data; and possibly, data from other federal agencies;
- TRACK-TBI and other relevant or Veteran-centric data sources.

Today we have eight DOE labs engaged. Through DOE's ESNet internet backbone, select VA sites can access the enclave to begin to work with the data. The physical and cyber security as well as privacy impact assessments have been an ongoing part of the stand-up of this effort.

The long term aim of MVP-CHAMPION today is to share DOE and VA resources in a reimbursable, fiscally-responsible manner that enables researchers to share expertise and thereby challenge and change current concepts in large-scale computing and health sciences. To guide the vision, VA and DOE teams need to work together not only on establishing the 'user facility'-like personal health information enclaves but as scientific collaborators. DOE and VA researchers need to understand each other's capabilities, challenges and mission needs in order to identify the high-impact research that will advance DOE and VA missions.

It will be important to also broaden the expertise and capability from the academic and commercial sectors to embrace the best of our country's technology base. There will be opportunities to bring new partners, data, expertise and capability to achieving our goals in aiding our veterans through efforts originating from DOE or VA. Additionally, the

VA/DOE team will need to pull from the best experts and resources across the full DOE national laboratory system and VA enterprise.

Goals

Simply put, the VA and DOE are working in cooperation to drive technology, innovation, and transform health care delivery for veterans bringing together an unparalleled and vast array of healthcare and genomic data with DOE's world class high performance computing (HPC), artificial intelligence and data analytics. By combining expertise, we are already pushing the frontiers of data analytics, next-generation computing, precision health, genomic sciences, and health care delivery. This partnership supports:

- Innovation tied to design and development of DOE's next generation supercomputing that will merge Big Data (BD), AI and High-Performance Computing (HPC) as well as innovation in population science using complex health system and genomic data for knowledge generation.
- Better Healthcare via using supercomputing to inform when and how to treat our veterans to improve outcomes and control cost.
- Better Science via a cadre of researchers and clinicians who specialize in healthcare with the DOE experts in HPC, AI & BD.
- Better Government via interagency collaborations bringing to bear the full capabilities and expertise within, and public private partnerships.

Priority Areas

On April 17-18, 2017, VA and DOE scientists, physicians, and leadership came together to develop technical roadmaps for how HPC can develop solutions to priority issues in caring for our veterans. VA priorities that were identified that could deliver early impacts were: Suicide Prevention, Prostate Cancer and Cardiovascular Disease. Specifically:

- ***Patient specific analysis for Suicide Prevention***: Suicide is the 10th leading cause of death in the US, and is significantly higher in the veteran population, with 20-22 deaths per day. Efforts would improve identification of patients at risk for suicide through new patient-specific algorithms built to securely provide tailored and dynamic suicide risk scores to bring the resources to each veteran at risk. Working closely with VA's Office of Suicide Prevention, the tools would be used to create a clinical decision support system that assists VA clinicians in suicide prevention efforts, and helps to evaluate the effectiveness of various prevention strategies.

- ***Help doctors discern lethal from non-lethal Prostate Cancer****:* Prostate cancer patients may undergo surgeries or other deleterious treatments without knowing if such treatments are actually necessary or effective, since there is no way of determining lethality *a priori*. The collaborative prostate cancer project will build classifiers for prostate cancer that could significantly aid doctors in distinguishing lethal from non-lethal prostate cancers. Reducing unnecessary treatments will

provide an increased quality of life for patients and allow the VA to focus resources where most effective.

- ***Enhanced prediction and diagnosis of Cardio-vascular Disease (CVD):*** CVD is the leading cause of death in US men and women - including veterans - and the cost of care for CVD conditions is high. A collaborative CVD project would build new predictive tools that (1) identify improved sets of risk factors for specific types of CVD and (2) develop methods that will inform individualized drug therapies. The new tools will enhance prediction, diagnosis and management of major CVD subtypes in veterans.

- ***Crosscutting Technology Advances:*** Cutting across the three projects are requirements for next generation AI and BD analytics. These requirements push development of DOE technologies in key areas including large scale data analytics, computer modeling, large scale machine learning and natural language processing. Success in developing these enabling technologies will have large impacts on DOE missions including science, energy, and national security.

Currently, the United States is the only country in the world with the opportunity to partner such a large scale health database with world leading data analytics capabilities. With this partnership, the US can drive cutting edge technology in next generation data analytics and improve quality of life for veterans and all Americans.

Future Partnerships

The FY 2019 VA Budget Request includes $27 million to support MVP-CHAMPION and a new program called ACTIV. This will deepen the collaboration between the VA and DOE and broaden the data and the potential impact. Secretary Perry and the VA Secretary have met many times to discuss these efforts, and we are now working on a broader MOA between our agencies. In May of 2017, Secretary Perry himself as a veteran joined the Million Veterans Program, donating his blood, his DNA and his records to the cause. The MVP dataset contains invaluable information on the connection between genes, environment, behavior and treatment.

As with the Human Genome Project, engagement and support from Congress was paramount to getting things off the ground and started in earnest. With funding from the VA and the expansion of the program these broader activities are potentially transformational for veterans' health and treatment. If we are successful, this will foster US innovation, help define and drive the intersection of artificial intelligence and big data analytics into the very fabric of DOE missions, transforming high performance computing from where it is today. It will define new standards in precision medicine and demonstrate the impact of scale in the application of HPC to medicine. When successful, veterans would be among the first to benefit.

Recently the VA and DOE held a meeting with Silicon Valley startups in computing precision medicine to understand the direction of the technology in the commercial

sector. As with the Human Genome Project, or the Exascale initiative today, partnerships with labs, academia and the private sector are important. We are still in our early stages of moving into this program due to its overall complexity, and we expect that partnerships will be essential.

Closing Remarks

We are starting to incorporate AI into our traditional computer simulations and exploring these new post-Moore design points. This ability to use data science and AI on the fastest supercomputers is a unique capability that depends on the hardware, software, and scientists and engineers at the National Laboratories. The scale of this approach is difficult to achieve any other way. Preliminary testing of an image analysis pipeline for traumatic brain injury magnetic resonance images took only 4 hours on a half-petaflop supercomputer with AI functionality, whereas traditional approaches take days to process an individual's brain scan. It is clear that there are few places where we can do this type of work.

Herein we have seen a natural way for both agencies to advance their missions, and this is informed by the ongoing VA-funded MVP-CHAMPION activities. The challenge of the VA dataset we believe will uniquely stress our computers, codes and people in dimensions that our existing datasets have not. On the other hand, the VA sees the benefit of access to a significant computing ecosystem that has extensive experience of working on complex problems at scale with frontier supercomputers. Additionally, our veteran's data is unique and must be treated with the utmost concern for privacy and other national security considerations. DOE provides a unique safe harbor for the data and applications of tools that is difficult to achieve in the commercial sector and with academic and commercial companies. Yet creating safe and secure opportunities for these stakeholders to work with the data and apply the cutting-edge tools developed in their laboratories and companies is essential for success.

At DOE, we see the challenge of predicting on these datasets as one of the most demanding of our time, therefore we have aligned our National Laboratories to stress the limits of frontier computing for hardware, software and our workforce in novel ways to inform the next generation of hardware acquisition; to inspire the next generation workforce to work on programs that impact the health and welfare of our country in new ways; to leverage the national laboratory system as it was designed to solve pressing problems of national significance that cannot be accomplished by individual investigators; and to innovate at a key moment where we have the unique ability to lead globally.

The United States is the only country in the world, at this moment, with the opportunity to partner a health database of this size with supercomputing resources and expertise. By partnering together, we can push the frontiers of computing and artificial intelligence, keeping DOE and the U.S. a world leader in science and technology. And by working together we can transform health research and healthcare for veterans and all Americans.