



Strategic Plan





3	Letter from the Director
5	Summary <ul style="list-style-type: none">Pillars of ActionFocus
8	Data Science at LLNL
11	Data Science in Action <ul style="list-style-type: none">Basic ScienceCognitive SimulationMaterials & Advanced ManufacturingNational SecurityPrecision Medicine
22	Implementation Strategy <ul style="list-style-type: none">ResearchEducationCommunityWorkforceFuture Vision
34	Tools for Collaboration <ul style="list-style-type: none">Open Data InitiativeComputational Resources
36	Administration <ul style="list-style-type: none">DSI Director and Administrative SupportData Science CouncilContactsAcknowledgments

DATA SCIENCE INSTITUTE

BIG

- Machines
- Data
- Ideas



LETTER FROM THE DIRECTOR



Building community, igniting innovative research, and providing strategic workforce development for LLNL's emerging and critical data science discipline

Large-scale data requires sophisticated techniques to interpret it, so data science has become an integral part of scientific research, particularly at Lawrence Livermore National Laboratory (LLNL). Awareness of this essential discipline has increased dramatically across LLNL. Programmatic work has benefitted from multimodal data analysis, data processing, outcome predictions, simulation, and experimental validation.

We founded the Data Science Institute (DSI) in early 2018 to facilitate mission-driven data science through a cohesive vision, increased collaboration, and targeted outreach and recruiting. The DSI has already made a significant impact on the data science community, both internally and externally, by building relationships between LLNL programs, staff, and academic partners. From the outset, we have endeavored to establish and promote LLNL's strategic vision in data science while combining collective expertise into a thriving community.

This coordination has roots in an earlier initiative. In 2013, LLNL launched the Data Science Initiative to align national security priorities with high-performance computing (HPC) capabilities and big data applications. The Initiative provided a focal point for advancing data science technologies for LLNL's unique scientific challenges. Over time, data science techniques became woven into LLNL programs, though the need remained to develop and sustain the workforce in this field. The Initiative transitioned into the Institute to adapt to the rapid pace of research and LLNL's growing portfolio.

The field of data science moves quickly. In a range of scientific research areas and through community-building objectives, the DSI is poised to advance LLNL's mission for years to come. I am honored to be part of this exciting initiative and to extend a warm welcome to current and future collaborators.

A handwritten signature in black ink, appearing to read 'M Goldman'.

Michael Goldman
Director, Data Science Institute

“LLNL is undertaking a significant effort to apply data science and machine learning techniques to scientific experiments in an attempt to better understand, predict, and control complex phenomena and systems. Through collaborations with multidisciplinary scientists, internal as well as external to the Lab, we hope to apply these techniques to test new hypotheses and enable novel scientific discoveries.”

— Anantha Krishnan
LLNL associate director of Engineering



SUMMARY

The DSI integrates with LLNL organizations to address the swift growth of data science and its impact on LLNL’s national security mission. The Institute is helping to cement LLNL’s data science identity and vision in national security through internal and external communication and coordination of activities designed to connect experts from LLNL, academia, industry, and other national laboratories. These efforts strengthen existing research, foster new collaborations, and grow and sustain LLNL’s data science workforce.

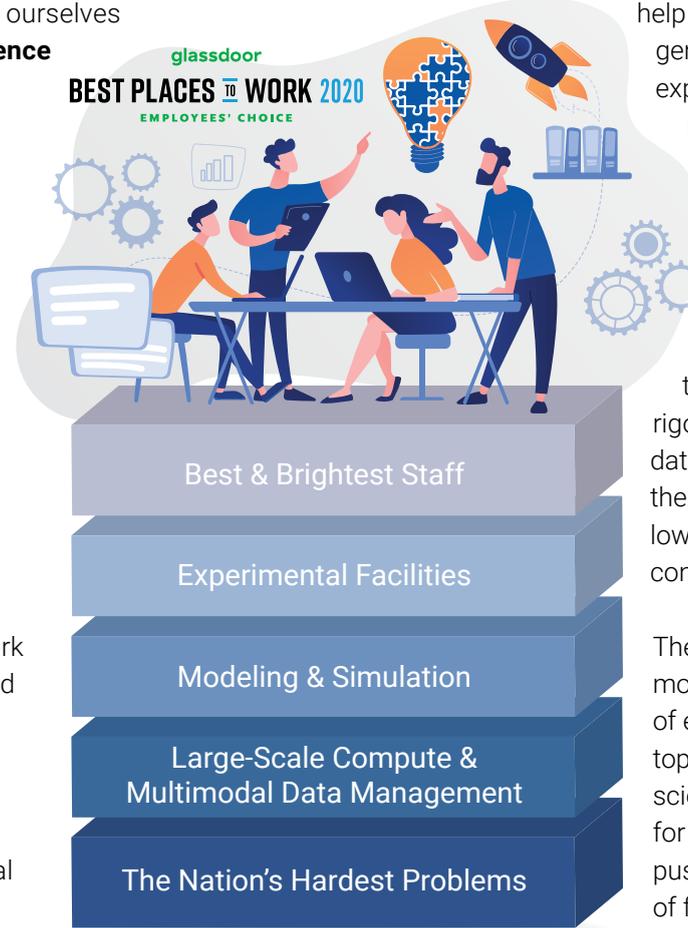
We promote and pride ourselves on **full-stack data science** to fully understand and deliver innovative solutions to some of the world’s most difficult and important challenges. This means the DSI and its community contribute to the entire hypothesis–experiment–analysis lifecycle, ultimately helping to accelerate scientific discovery.

Our data scientists work closely with first-in-field domain scientists to understand complex data and problems, drawing support from first-class experimental facilities, large-scale

simulations, and unique scientific datasets. Working with collaborators, we develop large-scale data management, collection, and ingestion techniques to explore unprecedented amounts of data. Through robust and efficient data architectures, subject matter experts and data scientists alike can rapidly and accurately analyze, assess, visualize, develop, interpret, and present new data and models.

We also invest in our data scientists professionally as leaders in their field. Our data science leaders help develop the next-generation workforce and expand diversity; influence research and development roadmap decisions through partnerships and technology development in line with the national interest; and ensure trust, accuracy, and a rigorous understanding of data analysis methods and their applications to meet low-risk tolerance for high-consequence missions.

These unified objectives will move us closer to our goal of establishing LLNL as a top-tier destination for data scientists with a passion for national security and pushing the boundaries of frontier science.



PILLARS OF ACTION

The Institute's strategic plan rests on four pillars of action that reflect the DSI's vision for shaping the future of data science at LLNL—collaborative research, outreach through novel education programs, development of a strong data science community, and workforce growth and sustainment—ultimately weaving data science into the fabric of LLNL's core disciplines.



FOCUS

Executing a five-year strategic plan will enable the Institute to achieve tangible metrics of success as we foster expanded collaborations in data science research, education, community, and workforce development. The DSI will invest in the following areas:



-  **Support leading-edge research**
 - Increased collaboration with industry and academic partners
 - Focused workshops to identify critical future research areas
 - Access to research findings and data that benefit a range of scientific areas
-  **Enhance coordination, communication, and awareness**
 - Community-building events hosted by the Institute
 - Expanded visibility of data science activities and research
 - Recognition as a leading entity in data science within our sponsor space
-  **Strengthen workforce**
 - Expanded internship opportunities for students, including growth of LLNL's Data Science Summer Institute (DSSI)
 - Novel educational programs in data science as related to LLNL mission
 - Increased availability of programs dedicated to developing internal staff into future data scientists

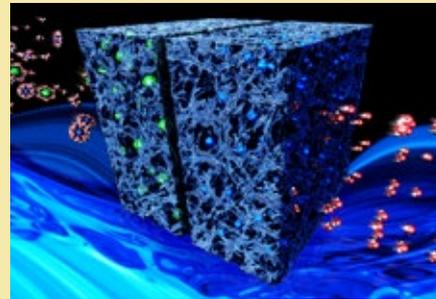
DATA SCIENCE AT LLNL

Big Machines. Big Data. Big Ideas.

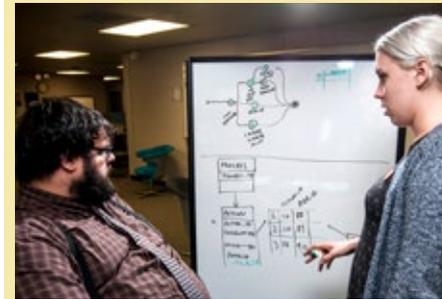
LLNL occupies a unique space in data science opportunities:



Home to some of the world's most powerful **supercomputers** and constantly evolving computing ecosystems.



A **data-rich environment** thanks to experimental facilities and complex systems.



A culture of **multidisciplinary teamwork** for solving some of the nation's most challenging problems.

Data science is a relatively nascent field, driven primarily by the large number of disciplines and applications that produce extensive amounts of complex data—data that can be used to develop predictive models to further scientific discovery. The DSI aims to provide intellectual leadership to help shape this new discipline by acting as the nucleus of data science activity at LLNL. The DSI works to (1) advance the state of the art in LLNL's data science capabilities to ensure the strength and robustness of the national security of the United States; and (2) help lead, build, and develop LLNL's data science community and workforce.

Understanding complex, interconnected systems

to enhance performance, resilience, and security is a national priority and a central mission for LLNL. The ability to capture ever-increasing amounts of data from these systems combined with advances in computational capability affords exciting opportunities in classic statistical and machine learning (ML) methods. This new paradigm has enabled the explosive growth of new predictive models that enhance our understanding of complex and critical systems.

The confluence of these advances has culminated in the formation of a newly focused, critical, and in-demand interdisciplinary domain known as data science. Data science brings an essential skillset

“LLNL is a data-rich environment. As the demand for sophisticated methods of analyzing and interpreting data grows, so too does the need to push the boundaries of data science.”

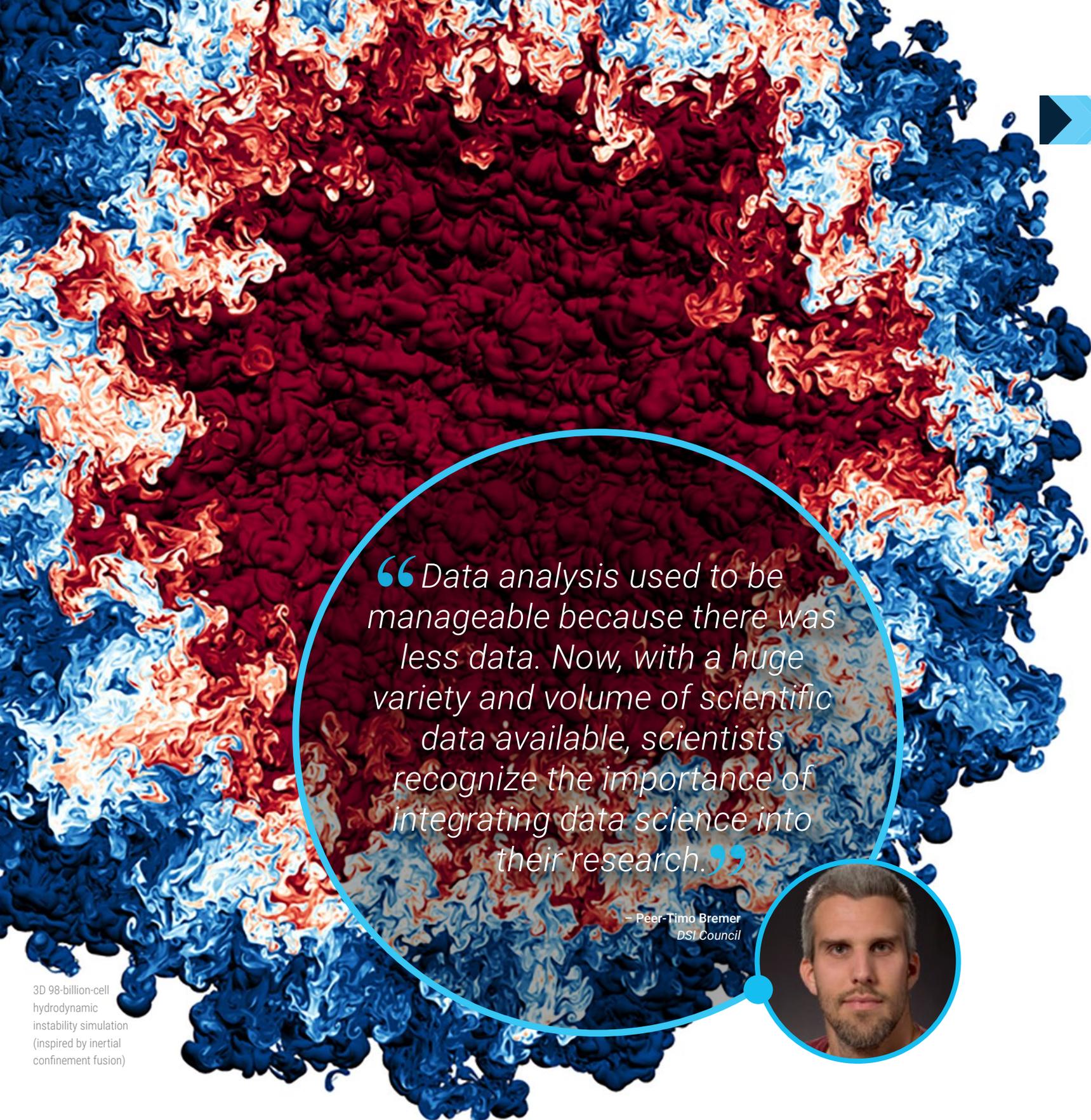
— Bruce Hendrickson
LLNL associate director of Computing



to the entirety of LLNL's mission with applications in biosecurity and human health, cybersecurity, advanced manufacturing and materials science, climate change resilience, energy security, nonproliferation, national and nuclear security, and space situational awareness, and the protection of U.S. critical infrastructure.

Given the high-consequence nature of these missions, LLNL also strives to be a leader in practicing and promoting responsible artificial intelligence to ensure data is appropriately safeguarded, inference agrees with scientific understanding, and applications align with LLNL's cultural and ethical values.





3D 98-billion-cell hydrodynamic instability simulation (inspired by inertial confinement fusion)

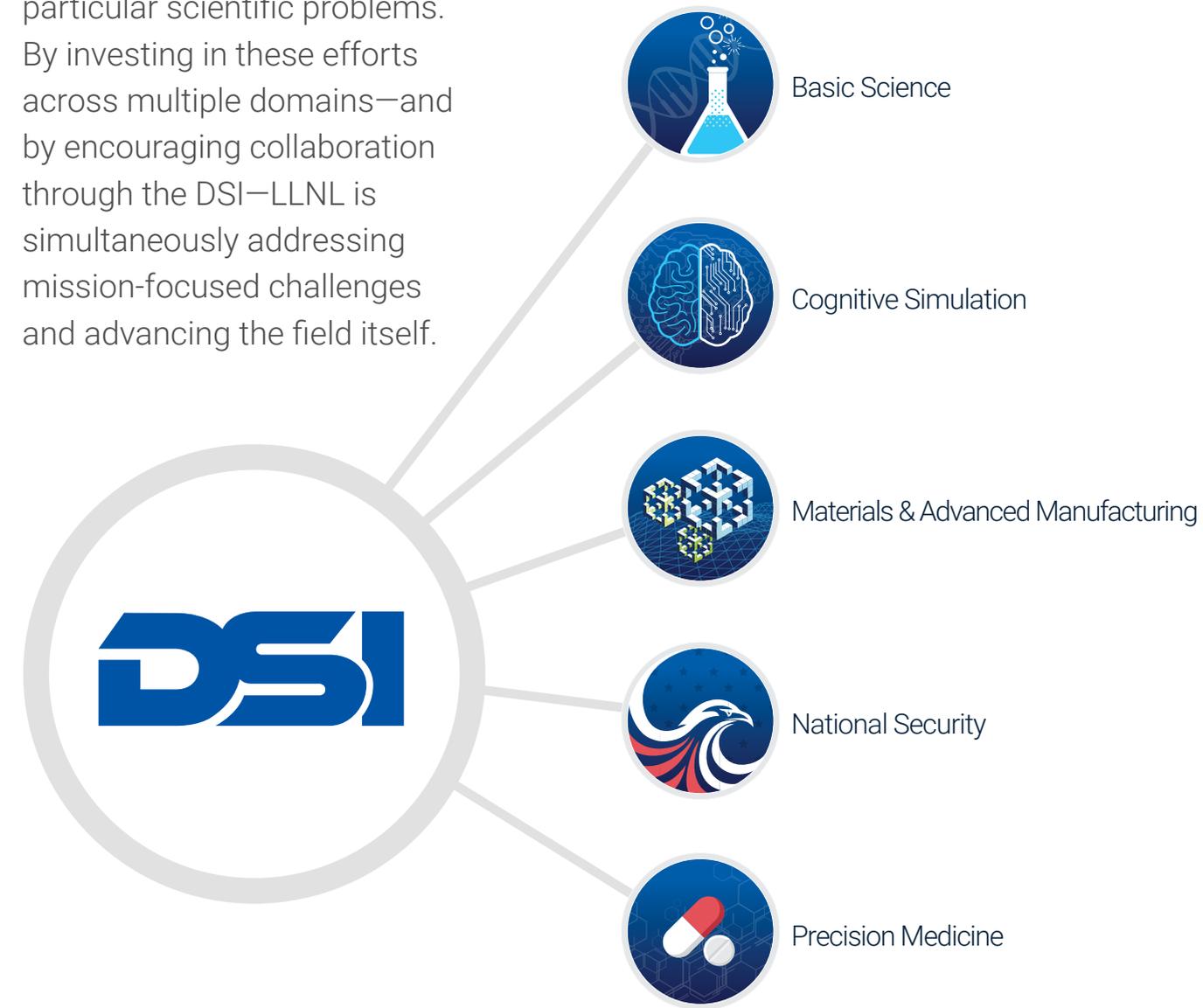
“Data analysis used to be manageable because there was less data. Now, with a huge variety and volume of scientific data available, scientists recognize the importance of integrating data science into their research.”

— Peer-Timo Bremer
DSI Council



DATA SCIENCE IN ACTION

The following examples of key LLNL programs rely on data science techniques to probe particular scientific problems. By investing in these efforts across multiple domains—and by encouraging collaboration through the DSI—LLNL is simultaneously addressing mission-focused challenges and advancing the field itself.





Basic Science

Analysis of telescope-gathered datasets leads to greater insights about the universe.

Astrophysics is a major growth area in LLNL's advancement of basic science for national and global security needs, and data analysis is crucial when it comes to probing big questions about the physical universe. Our scientists strive to describe fundamental characteristics of the Solar System to push the frontiers of our physical understanding of the universe.

With modern telescopes and detectors, astronomy has transitioned from the study of individual objects to statistical characterizations of populations throughout the history and expansion of the universe. Big data computing pipelines, precision modeling, and scalable ML are the tools required to make sense of **modern astronomical data**.

Multiple astrophysics projects at LLNL explore questions around the nature of dark matter, dark energy, and space situational awareness. For example, our scientists study the nature of dark energy by developing image-analysis algorithms that make inferences from low signal-to-noise data. This process requires evaluation

of **statistical versus systematic errors** in the dataset. Whereas statistical errors shrink as the dataset grows, systematic errors are caused by variables that do not simply average out—for instance, atmospheric turbulence or imperfections in a camera lens. Researchers account for these errors with hierarchical Bayesian forward models, simulations of noise, and data-processing software.

One major strategic initiative combines Bayesian statistical methods and deep learning (DL) to catalog and interpret debris, satellites, and other objects orbiting Earth. Exploiting a relatively small-scale dataset, this effort aims to develop an **advanced predictive physics model** to evaluate less data with higher accuracy. Similar techniques could enable collaborative autonomy within a constellation of satellites tasked with tracking orbiting objects. In this setup, each satellite exchanges data with the rest of the network until they reach consensus on an object's location. This capability could help predict when space debris will hit a satellite or Earth.

Modern astrophysics relies on statistical characterizations of historical and large datasets.

Dark matter and dark energy studies are among the investigations benefitting from ML algorithms.

Space situational awareness projects use ML and statistical methods to track potentially dangerous orbiting objects.



Cognitive Simulation

ML integrated with simulations can effectively and accurately compare large-scale simulations and experiments.

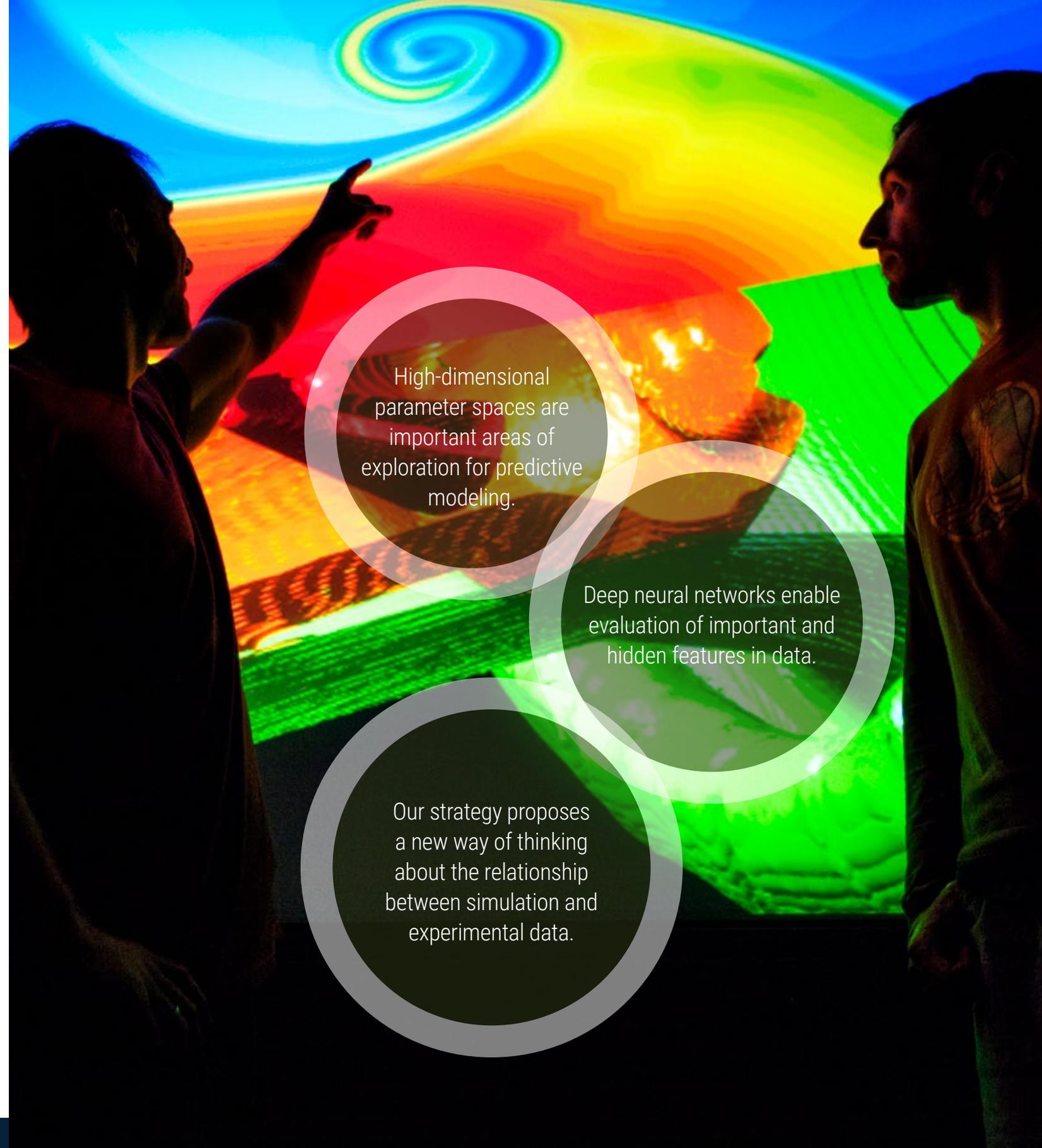
A central objective of LLNL's national security mission is to improve and advance **predictive simulations** by challenging them with precision experiments. This integration of simulated and experimental data is increasingly driven by large-scale data analytics. Modern research that addresses core challenges—such as maintenance of the U.S. nuclear stockpile, nuclear nonproliferation, pharmaceutical design and cancer research, engineering design optimization, and the resilience and security of critical infrastructure—relies increasingly on HPC simulations and experiments that produce data of **unprecedented complexity**.

Contemporary modeling efforts must incorporate and balance high-dimensional model parameter spaces, multiscale and multifidelity analysis, uncertainty quantification (UQ), and robust experimental data. Researchers must compare simulation output with experimental data to adapt and improve predictions. However, the scale and complexity of both experimental and simulated data has moved beyond that which humans can handle in their heads. To accomplish this, LLNL has begun a new strategic initiative—called cognitive simulation—that uses ML to help scientists navigate this new data-rich environment to improve simulation models.

The backbone of the project's cognitive simulation approach is leveraging **deep neural networks** to reveal structure in large datasets by learning to map experimental inputs to simulation outputs. This work involves engineering and exploiting latent spaces, which lie at the heart of ML processing and enable researchers to evaluate important—and even hidden—features of compressed data.

The team is **integrating ML into simulation** in four ways:

1. **"in the loop"** algorithm or resolution switching, in which approximate regression models replace complex physics calculations;
2. **"on the loop"** prediction and correction of mesh tangling and step-wise simulation execution;
3. **"steering the loop"** learning that proposes the next simulation needed to reach an optimal dataset and stops uninformative ensemble members from continuing; and
4. **"after the loop"** improving ML models by incorporating experimental data.



High-dimensional parameter spaces are important areas of exploration for predictive modeling.

Deep neural networks enable evaluation of important and hidden features in data.

Our strategy proposes a new way of thinking about the relationship between simulation and experimental data.



Materials & Advanced Manufacturing

Data science advances researchers' ability to understand, develop, and deploy new materials.

LLNL's development of innovative materials ranges from advanced metallic nanowires to high-performance alloys, from freestanding polymer films to components that refresh the aging nuclear stockpile, and much more. Regardless of the final product, ideal feedstock materials need to be synthesized and optimized for system-level integration that will meet performance requirements with predictive behaviors. However, the **materials discovery process** can take 10 to 15 years before application integration.

A series of projects aim to accelerate the materials discovery, optimization, and deployment processes by using data science techniques. In one innovative study, a team was inspired by the growing volume of scientific literature to **extract targeted information** from published papers—thus automating the repetitive, formidable task of looking for new protocols and chemicals used for creating materials. An LLNL-developed tool contains text from tens of thousands of papers and allows the user to analyze data for different variables. The extraction pipeline begins with a supervised logistic regression algorithm that highlights recipe-like sentences. Another

algorithm combines a conditional random field model with natural-language processing to extract chemical information. Visualizations render the data for further analysis.

Another project goes beyond text descriptions of materials to analyzing images of other materials whose physical properties often correlate with performance but are difficult to quantify. A **feature-extraction tool** leverages open-source technologies that define engineered features, such as boundary detection, at a pixel level. The computer learns to weigh feature importance, then provides computed values that translate to mechanical performance prediction.

Data-driven techniques further enhance the development of specialized materials by combining multimodal data for feature extraction. With the help of data visualizations manipulated in a custom-built user interface, material properties are correlated with high-explosive performance via identification of features in images and numerical values from varied sources. This project advances the application of ML algorithms for small datasets while also implementing physics-based approaches.

Data-driven analysis techniques accelerate materials optimization and development efforts.

Researchers extract chemical information, physical properties, and other features from datasets.

Key efforts are aimed at improving high-explosive materials performance.



National Security

Multimodal data analytics advance nuclear proliferation detection methods.

Nuclear nonproliferation is of paramount concern to the Department of Energy, the National Nuclear Security Administration (NNSA), and other government agencies. LLNL's national security mission includes developing scientific and technological solutions to address the evolving landscape of proliferation threats. This means monitoring and detecting weapons of mass destruction as well as preventing the spread and availability of related materials and infrastructure. **Early detection** of proliferation activity is one key LLNL effort in this area that relies on data science techniques.

Analysts reviewing and interpreting proliferation activity data face numerous challenges: Beyond contending with an increasing volume, rate, and variety of information, they encounter unstructured, improperly labeled, or even unlabeled data. Moreover, a dataset may have a low density of highly valuable data—in other words, a needle in a haystack. Analysts must quickly identify specific, nuanced nuclear proliferation processes under these circumstances.

A multidisciplinary LLNL team is developing **DL algorithms** that map images, video, text, and audio into a joint semantic feature space that conceptually relates different forms of data to the analyst's query. Nicknamed the "Semantic Wheel," the technique leverages new scalable training algorithms that take advantage of LLNL's world-class supercomputers for rapidly training large neural networks on massive datasets. Relationally, each type of data is a spoke in the wheel, uniting with other spokes in the feature space at the wheel's center.

The Semantic Wheel relies on a two-pronged ML approach. First, the team develops feature-learning algorithms that are self-supervised and unimodal—individual imagery, video, text, or audio modality—and therefore enable learning of high-quality, transferable representations from large, unlabeled datasets. Then, multimodal learning algorithms merge the unimodal representations into a shared semantic feature space. This work to **interpret multimodal datasets** is significant for the future of proliferation detection.

Nuclear proliferation activities are extremely difficult to detect.

Innovative ML algorithms help analysts identify and interpret multimodal data.

This work established a foundation for a multi-institutional Advanced Data Analytics for Proliferation Detection project.



Precision Medicine

Data-driven analysis techniques can reduce the timeline of the cancer treatment pipeline.

The amount and value of healthcare-focused, experimental molecular measurements are growing rapidly. Access to this data presents new opportunities to learn more about the molecular drivers of disease and to develop improved options for treatment. In cancer research, **data-driven molecular analysis** helps scientists predict response to drugs.

For example, to develop a new drug that safely and effectively targets cancer cells or the processes that cause cancer, researchers typically rely on costly, time-consuming, and high-failure experimentation with different combinations of molecules. The multi-institutional Accelerating Therapeutics for Opportunities in Medicine consortium—which includes LLNL experts in modeling, simulation, and data science—is speeding up the **drug discovery pipeline** by leveraging specialized HPC hardware and software to accelerate molecular analysis and scale predictive ML algorithms. With these combined resources, the consortium aims to transform drug discovery into a rapid, integrated, and patient-focused process—and thus achieve better outcomes.

The consortium’s “chemistry design loop” approach optimizes the processes of exploring chemicals, rapidly evaluating molecules, and proposing new compounds. Among other activities, the team is building a library of ML models for automated comparison of hundreds of drug properties known to be important for drug design. This evaluation relies on a sophisticated ML pipeline that integrates traditional statistical models with DL algorithms. The process generates **predictions of new simulations** to run and experiments to conduct by ranking top molecules as well as those with less confident predictions. Prediction quality is expected to improve as the algorithms learn from previous iterations, starting the loop anew.

Additionally, the team works toward understanding the criteria needed to make accurate ML predictions on candidate compounds and, crucially, when they will fail. A key goal is ensuring that new data collected will be valuable in further algorithm training. **UQ analysis** guides active learning, characterizes confidence in model predictions, and assigns weight to model ensembles such as random forests and neural networks.



An innovative consortium combines multiple disciplines to accelerate cancer drug discovery.

ML techniques enable automated comparison of hundreds of drug properties.

Prediction quality for candidate compounds increases with iterative learning and UQ.

IMPLEMENTATION STRATEGY



Supporting the growing dependence of data science to LLNL's mission, the DSI builds relationships between LLNL, academia, and industry to strengthen workforce, education, and research. To be effective in its role, the Institute has developed a five-year implementation strategy built on its four pillars described above: research, education, workforce, and community.



The Institute's pillars focus on expanding collaborative research across a broad range of LLNL mission areas, growing the workforce to support new research

through internships and other student programs, strengthening the workforce through a variety of DSI and partner learning opportunities, and

fostering a sense of community to sustain this critical workforce.

THE FIRST YEAR of our strategic plan will focus on establishing pilot programs within each pillar, identifying potential collaborations to help advance our strategy, and refining future year strategies. Programs deemed successful will continue and expand in later years.

YEARS 2-4 are evolutionary years and will focus on expanding successful pilot programs and increasing the DSI's reach in the data science community.

YEAR 5 will be used to reflect on lessons learned and develop plans to move the Institute forward into future years.

YEAR
1



- Facilitate new research and educational partnerships
- Host a seminar series with invited speakers
- Co-host a workshop in data science with the University of California (UC)
- Launch pilot education and community programs
- Formalize the student internship program

YEARS
2-4



- Expand student internship program and form new student, postdoctoral, and faculty opportunities
- Formalize academic partnerships, specifically focused on leveraging relationships within the UC system
- Establish new partnerships with industry collaborators to develop next-generation compute architectures and algorithms that are optimized for large-scale ML applications
- Evolve annual workshop into micro-workshop focused on establishing new research trajectories in data science
- Scale successful education and community programs
- Highlight and open access to select LLNL datasets with pilot data- and software-sharing programs, available to internal and external researchers alike

YEAR
5



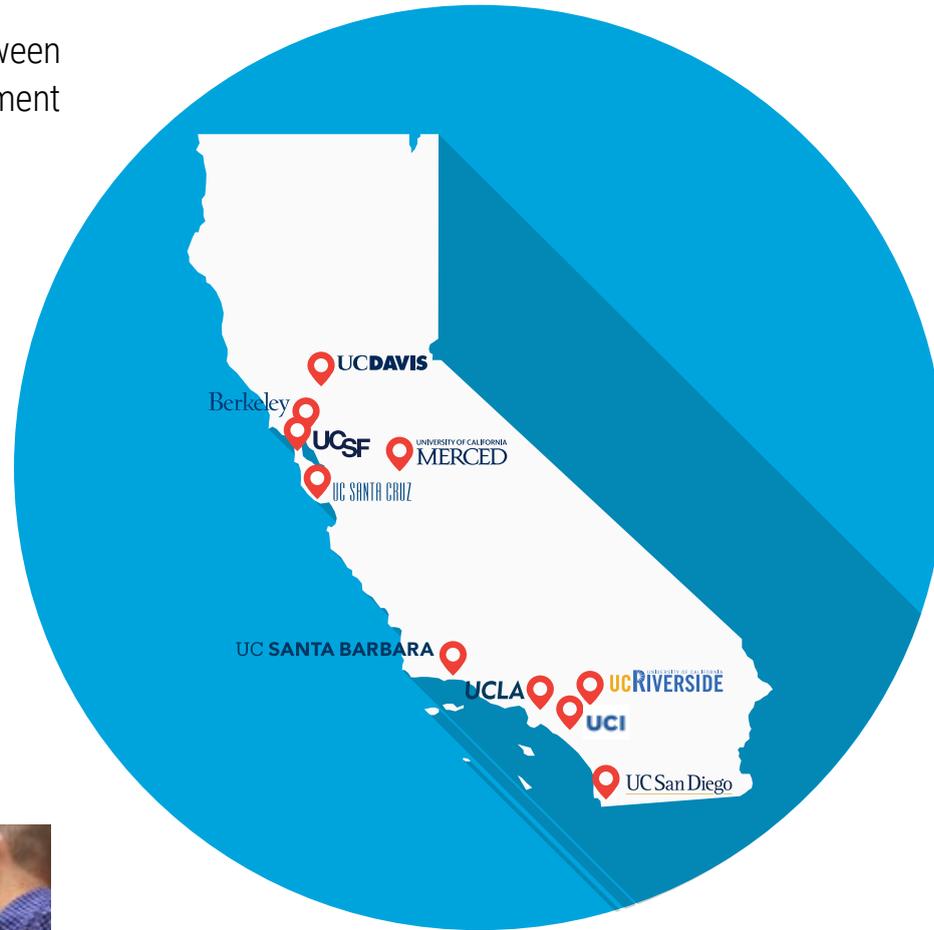
- Expand academic and industry collaborations
- Cement workforce pipelines from UC to provide strong, diverse data science candidates to LLNL
- Develop formal education plan with UC to democratize data science education to LLNL workforce and promote national lab mission to new undergraduate and graduate students
- Establish a permanent repository for LLNL datasets and software to foster further collaboration, education, and research activities
- Complete plans for the DSI's future direction



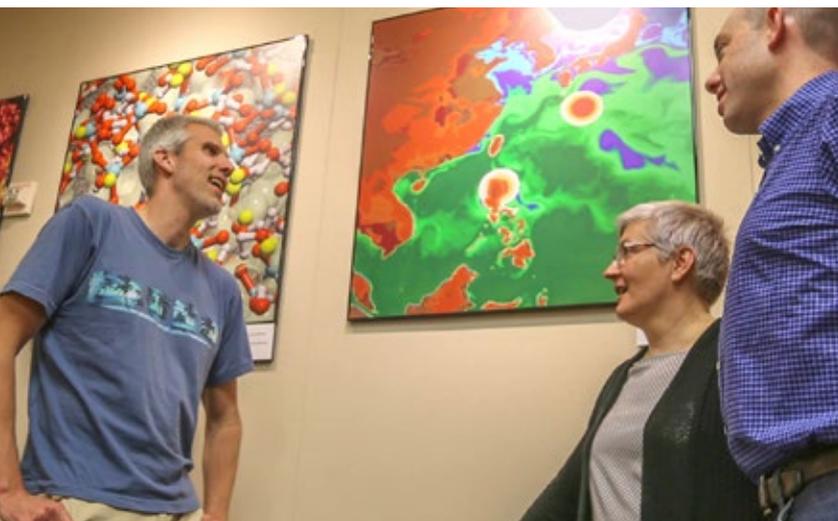
RESEARCH

Pushing the boundaries of scientific study through multidisciplinary collaboration

The DSI facilitates vital collaborations between academia, industry, and other U.S. government agencies—thus evolving LLNL’s strategic data science vision and guiding future science and technology investments. These collaborations will energize LLNL staff; stimulate new thinking; yield opportunities to mentor young scientists; and increase awareness of LLNL’s mission, research, and capabilities. Such activities are essential to maintaining LLNL’s place at the forefront of innovation and keeping an engaged workforce.



The DSI is primarily engaged with the UC system and is actively focused on continuing existing partnerships and building close collaborations with UC campuses.



Year 1:

The DSI launched several initiatives to enrich cross-domain research collaborations and increase awareness of LLNL research.

Highlights include:

- The inaugural **Data Science Workshop** invited more than 200 LLNL staff, UC students and faculty, and other UC national labs to participate in a multi-day event to share and discuss a broad range of data science topics across a wide set of application areas. This event established a multitude of new connections among internal staff, academic and laboratory researchers, and students.
- The **DSI Seminar Series** hosted a dozen speakers from academic institutions and industry in its first year, introducing new ideas and connecting potential collaborators to LLNL. Seminars ranged from in-depth discussion of methods, such as indirect supervision and deep convolutional networks, to applications in brain tumor imaging, genetic engineering, and even the social sciences.
- The **DSI Consulting Service** was established to provide free, short-term resources to connect data science experts with LLNL programs across all domains looking to leverage advanced data science techniques to develop novel solutions.
- The **Open Data Initiative (ODI)** was launched to share with the data science community at large LLNL’s rich, challenging, and unique datasets. These datasets will help support curriculum development, raise awareness around LLNL’s data science efforts, foster new collaborations, and be leveraged across other learning opportunities.

Years 2–4:

The DSI will take the following steps to support this pillar:

- Expand our collaborations with UC to include all 10 campuses and expand our strategic partnerships beyond the UC system to include the NNSA, U.S. government, and other top-tier academic institutions broadening the community’s awareness of LLNL’s unique research and applications in data science.
- Continue to architect new and foster existing initiatives and programs to help expand collaborative research—such as new faculty and fellow programs, micro-workshops focused on selected topics of interest from the DSI’s annual workshop, support of internal research and development opportunities, computational testbeds, and resources to support future research.
- Establish partnerships with UC campuses to leverage existing data-hosting infrastructure (such as the California Digital Library) to make LLNL datasets, as identified by the ODI, more widely available to academic researchers across the UC system and beyond.

“Data science presents an enormous opportunity to advance the Lab’s missions, providing new ways to utilize a broad array of data and our increasingly sophisticated simulation capabilities. The DSI is an important bridge for the Lab to the broader data science community to speed the learning process.”



– Kimberly Budil
Principal associate director of Weapons and Complex Integration



EDUCATION

Creating learning opportunities and strengthening skills in a fast-moving field

The DSI partners with UC campuses to cultivate educational programs, curricula, and datasets that generate an array of data science learning opportunities. These objectives align with LLNL mission areas and keep pace with the fast-moving field. For example, we envision partnering with formal degree programs that allow students to gain experience working with LLNL problems and staff while earning their degrees and garnering an understanding of LLNL's mission and challenges. Simultaneously, LLNL staff may leverage these same resources to learn new and strengthen existing skills.



Year 1:

The following activities supported this pillar:

- Development of a **data science education plan** to allow LLNL staff to self-select skill paths that are relevant to supporting their programs or support their desire to change their career path. This education plan allows staff and other collaborators to quickly and easily identify important skills for a variety of application areas and provides educational resources to formally or informally acquire those skills.
- Formation of **reading groups** for LLNL staff, students, or research collaborators to identify and discuss the state-of-the-art research in a variety of data science topics—such as natural language processing, deep neural networks, statistical modeling and analysis, and deep reinforcement learning.
- Development of a pilot **data science challenge** program where DSSI staff partnered with UC Merced faculty to teach a cohort of undergraduate and graduate students the skills required to develop solutions to challenging data analysis problems. LLNL provides datasets, subject matter experts, and lectures over the course of an intensive, hands-on two-week period.
- Introduction of a **data science immersion** pilot program to provide a year-long internship comprised of multiple projects and mentors. Through this program, qualified nontechnical staff gain on-the-job training and experience as a data scientist to facilitate career growth and strengthen LLNL's core data science workforce.
- **Visiting faculty summer programs** to promote collaboration and provide short courses on selected topics of interest to LLNL staff and students. Participating faculty home campuses include UC Santa Cruz, UC Merced, UC Riverside, Virginia Tech, and Brigham Young University.

Years 2–4:

The DSI will expand these efforts to strengthen the existing LLNL workforce as well as promote and support learning opportunities with a broader set of collaborators, helping to educate future generations of data scientists through the following activities:

- Implement the education plan through an interactive, web interface to crowdsource feedback and track new educational resources.
- Formalize the immersion program and open the opportunity to a larger subset of LLNL staff.
- Work with the UC system towards a national laboratory-centric data science curriculum and formal degree program.
- Promote and leverage sabbatical programs and other visiting faculty programs to expand collaborative research and increase availability of faculty-taught courses to LLNL staff.
- Establish new educational partnerships with select universities to enable LLNL staff to teach for-credit courses at undergraduate and graduate levels.



“Given its access to the leading-edge science at LLNL, our work with the DSI to provide meaningful fellowships for graduate students has allowed us to grow and strengthen our data science workforce in ways we would not have on our own.”

– David Mongeau
Executive director of the Berkeley Institute for Data Science



COMMUNITY

Engaging researchers through strategic outreach and partnerships

The DSI serves as a communication switchboard between the Laboratory, academia, industry, and other national laboratories to ensure those communities are engaged and aware of the latest research, existing and potential partnerships, and workforce needs and opportunities. Accordingly, the Institute will facilitate cross-pollination across communities to form new research areas both within LLNL and externally.



“Collaborating across domains with similar data analysis needs is crucial for strengthening the networking and educational opportunities within the data science community.”

– Marisa Torres
LLNL Women in Data Science
ambassador

Year 1:

The Institute has been promoting and expanding visibility of LLNL data science activities through:

- On the website, the **Data Scientist of the Month** program highlights a new LLNL data scientist every month for exceptional work or contributions in the area of data science.
- Increased **outreach initiatives** establish a technical and information presence at top-tier, highly visible conferences as well as at nontraditional conferences, workshops, and other events where there is a data science presence.
- The **latest LLNL research**—including conferences, papers, and other publications—is tracked and publicly

highlighted on the DSI website.

- **Monthly bull sessions** provide a forum for LLNL staff, collaborators, and students to discuss technical or workforce challenges, identify new research areas, or share information about newly discovered resources.
- **Annual town hall meetings** are held to hear from LLNL students and staff about ways the DSI can better support data science and its workforce across LLNL.

In addition to these specific activities to foster a community at LLNL, previously highlighted activities—workshops, seminar series, academic partnerships, the ODI, and summer and faculty programs—are all instrumental in connecting

with external communities to keep up with the latest research, partnerships, and challenges in data science. This effort includes outreach and sponsored sessions at premier events that promote diversity in scientific study—for example, the Women in Data Science Conference, the Women in Statistics and Data Science Conference, the Conference on Knowledge Discovery and Data Science, and the Grace Hopper Celebration.

Finally, the Institute is responsible for the majority of data science-related communication across LLNL, acting as the window into data science efforts, staff, and needs for all Laboratory organizations. This is accomplished through the DSI website, which is updated frequently with new informational content, job opportunities, and other information for those interested in data science and related activities at LLNL.

Years 2–4:

DSI will continue to expand community-building activities to enhance and sustain workforce strength through the following activities:

- Record and broadcast the seminar series to allow for remote participation or later viewing by LLNL staff and the larger data science community.
- Continue to increase outreach activities through informational sessions, student club interactions, datathons, and workshops.
- Facilitate more frequent and productive discussions and activities by co-locating available LLNL staff in the Livermore Valley Open Campus (LVOC) and providing convenient access to academic and industrial partners, students, and researchers.





WORKFORCE

Developing and guiding the next generation of data scientists

As data science is increasingly relevant to LLNL's mission, it is critical for the DSI to develop, grow, and sustain a successful workforce pipeline. The DSSI was established to address this need by recruiting diverse and talented data scientists from the academic community into LLNL.

“DSSI students have made significant contributions to their projects, and they form a great pool of candidates for technical staff positions at the Laboratory.”

– Goran Konjevod
DSSI director



The DSSI is:

1. A flexible 12-week summer internship program for graduate and undergraduate students in data science and related fields.
2. A central pipeline for recruiting data science staff into LLNL.
3. Responsible for providing relevant experience to students. Former DSSI interns have received job offers with LLNL, other national laboratories, and industry. The program has also encouraged many undergraduate students to pursue advanced degrees.
4. An opportunity to foster the creation of academic partnerships through coordinated activities and projects for both students and faculty.
5. Led by director Goran Konjevod and co-director Marisol Gamboa.

“The data we give the interns is not synthetic. The scenarios are not merely hypothetical. Students have a huge opportunity to explore cutting-edge science here.”



– Marisol Gamboa
DSSI co-director

ORGANIZATION:

- Students are selected through an open competitive process and paired with LLNL mentors and projects that align with their skills and interests.
- Students are co-located onsite to foster collaboration, centralize activities, and create a fun and stimulating environment.
- Students are supported by projects with their mentors and are provided up to 50% of their time to explore different projects, participate in other Laboratory activities, attend courses and seminars, and strengthen their skills in data science.

ACTIVITIES:

- Short courses and mentorships are provided by selected visiting faculty.
- Additional courses are offered by LLNL staff.
- Challenge problems leverage large and varied datasets used and/or created from actual LLNL projects. Students work in teams on challenging questions involving real-life data that represent the breadth and depth of data science work at LLNL.
- Coordinated discussions connect students with senior technical and workforce leaders.
- Students develop their own activities such as social groups for extracurriculars or technical “un-seminars.”

FUTURE:

- Scale up the summer program year-round, supporting academic fellows, faculty, and coop or externship students.
- Scale up the program through increased UC participation to expand UC student and faculty representation, including continuing outreach to first-generation college students and those from underrepresented populations.
- Work with NNSA consortia to partner with student programs and faculties to enhance cross-pollination between subject matter experts across the natural science fields and data science experts.
- Leverage the DSI's ODI to supply representative challenge problems and datasets.
- Leverage advances in workflow and development tools to give students easier access to data science technology stacks and computational resources.

FUTURE VISION



“Our missions demand that we have people working to create and apply the most advanced tools in the challenging and essential field of data science. There is an exciting future ahead for data science research, development, and application at LLNL.”



– Pat Falcone
LLNL deputy director for
Science and Technology

During the fifth year, the DSI will deliver a retrospective report that includes recommendations regarding a future direction for the Institute. The report will describe our analysis of each initiative, provide an assessment of what is working and what did not, and outline remaining challenges that need to be addressed. The report will also be delivered to a community of LLNL and academic stakeholders.

At the end of five years, we expect that the Institute will have achieved the following milestones:

- Strong internal and external community support demonstrated by continued and growing sets of activities revolving around data science, newly identified research areas and partnerships, and consistent workforce engagement.

- Evidence that the workforce pipeline is satisfying LLNL's demand for data scientists with diverse, high-quality talent coming out of universities.
- A recognized brand establishing LLNL as a top-tier destination for data science research and applications.
- Partner and public access to curated LLNL computational resources and datasets that necessitate large-scale computation through the DSI ODI.
- An education plan for remote learning opportunities and curricula.
- Addition of leading industry and academic data science experts to the DSI's external advisory board.

In addition, we will explore opportunities to establish a physical presence in the form of an external hub at LLNL's LVOC. The hub will potentially leverage the UC facilities present there as part of our planned expansion of collaboration with UC faculty and students. LVOC will also provide postdocs, students, faculty, and LLNL researchers a dedicated, common, and accessible space that will facilitate collaborative interaction through the use of additional office, meeting, and networking space. It will further enable, and in some cases be crucial to, many activities outlined in this strategic plan such as remote learning programs and expansion of student

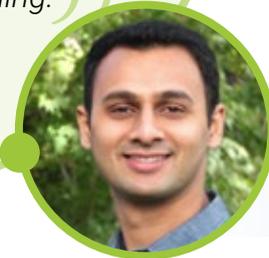
programs. LVOC will provide convenient and modern space for seminar speakers and workshops.

As an expanding village of advanced scientific and engineering centers, LVOC is a novel venue for collaboration with LLNL. Researchers from private industry and academia partner freely with Laboratory personnel to contribute to the next generation of big ideas. LVOC will play an important role in promoting collaborations and strengthening the multidisciplinary teamwork that is crucial to advancing data science as a field and the scientific applications that rely on it.

TOOLS FOR COLLABORATION

“The ODI is a Laboratory-wide effort to make our rich data ecosystem available to the broader data science community. Open datasets have been a crucial factor in the past decade’s progress in machine learning. Our open datasets will help drive the next decade of advances while addressing unique challenges in scientific machine learning.”

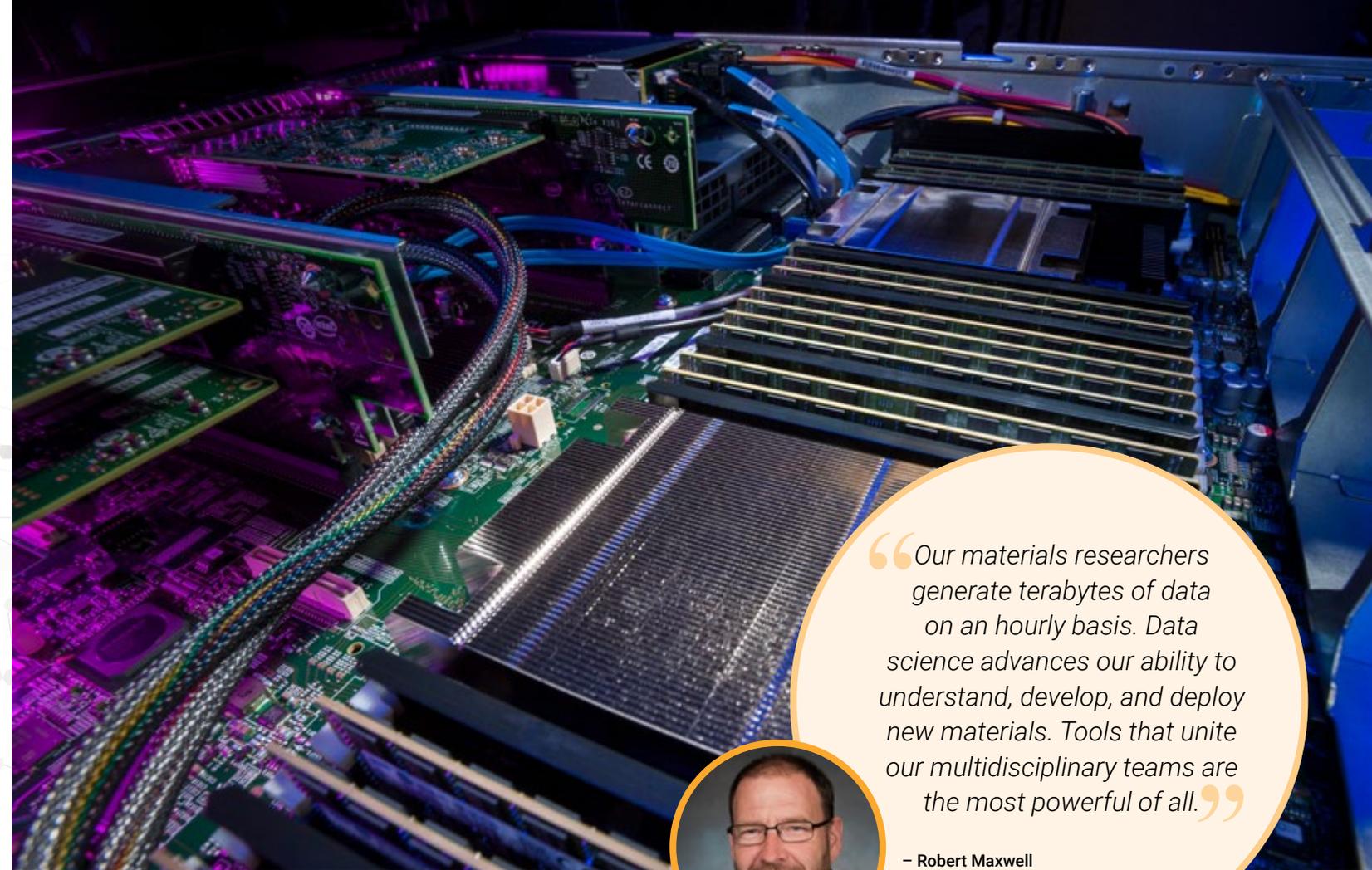
– Rushil Anirudh
ODI director



Open Data Initiative

The DSI is working to release a large number of unique LLNL datasets for public use as well as maintain limited-release datasets for targeted research partnerships. These ODI datasets will help drive interest and engagement across LLNL staff and academic researchers, students, and postdocs to develop novel solutions to some of the nation’s hardest problems. These datasets will represent a wide variety of key LLNL mission areas; will be representative subsets of some of the world’s largest datasets; ranging in complexity from dense, featureful, labeled datasets with well understood solutions to those that are sparse, noisy, and largely unexplored. Access and research activities include:

- Academic partners will get a “first look” to support and inspire newly developed undergraduate and graduate curricula.
- LLNL students, postdocs, and visiting faculty will have unfettered access to large, limited-release datasets.
- LLNL staff will have easy access to help them explore applications of algorithms, models, techniques, and methodologies across domains to identify novel solutions and develop new research areas.
- We will strive to form new academic and industry partnerships as future collaborators demonstrate novel solutions on ODI datasets.
- We will leverage ODI datasets to test novel hardware solutions for future scalable ML platforms.



“Our materials researchers generate terabytes of data on an hourly basis. Data science advances our ability to understand, develop, and deploy new materials. Tools that unite our multidisciplinary teams are the most powerful of all.”

– Robert Maxwell
LLNL Materials Science Division leader



Computational Resources

LLNL is home to some of the world’s most powerful HPC systems, which support our large-scale experimental facilities and enable scientists to perform new algorithm and model development and analysis at unprecedented scales. With constantly evolving hardware and software resources and state-of-the-art infrastructure, this world-class computing environment also offers a wealth of expertise in building, porting, running, and tuning real-world, large-scale applications. Together, these unique compute platforms and LLNL’s expertise will:

- Enable scientific advances in predictive models spanning multiple domains including materials science, precision healthcare and biosecurity, energy, climate science, cybersecurity, and other areas of national security.
- Serve as testbeds to integrate, test, and develop novel hardware and software solutions for the scalable ML and statistical analysis required by increasingly complex experimental data.
- Provide collaborators with account access, hardware and software resources, data management, and training.

ADMINISTRATION

DSI Director and Administrative Support

Michael Goldman serves as the Institute's director and is responsible for establishing the DSI's mission and strategy while executing its day-to-day operations.

Jennifer Bellig serves as the DSI's administrative support and central coordinator.

Data Science Council

The Council is instrumental in consulting on LLNL's overall data science strategy and helping to execute DSI activities. Their dedication and leadership contributed to the development of this document and the ongoing success of the DSI. Members include Peer-Timo Bremer, Barry Chen, Daniel Faissol, Ana Kupresanin, and Michael Schneider.

The DSI extends special thanks to David Buttler, Cassandra Fronczyk, and Katie Schmidt for their hard work as ambassadors and their contributions to this document and overall DSI activities. Cindy Gonzales and Emily Brannan were indispensable during the DSI's inaugural year with administrative support.



Contacts

Michael Goldman | goldman21@llnl.gov | 925.423.9422

Jennifer Bellig | bellig1@llnl.gov | 925.424.5197

data-science.llnl.gov

Acknowledgments

The Council would like to acknowledge the many supporters of the DSI for their advice and guidance in the preparation of this document, including LLNL staff Holly Auten, Emily Brannan, and Mary Gines. External advisors were David Mongeau (UC Berkeley Institute of Data Science), Pamela Reynolds (UC Davis), Abel Rodriguez (UC Santa Cruz), Bruno Sanso (UC Santa Cruz), Harold Smith (UC Berkeley), and June Yu (UC Office of the President).

©2020

Lawrence Livermore National Laboratory is operated by Lawrence Livermore National Security, LLC, for the U.S. Department of Energy, National Nuclear Security Administration under Contract DE-AC52-07NA27344. LLNL-AR-801962

