SUMMARY REPORT JULY 2021 – JUNE 2023

GROWING A VERSATILE COMPUTING ECOSYSTEM

SDSC

SAN DIEGO SUPERCOMPUTER CENTER

UNIVERSITY OF CALIFORNIA SAN DIEGO

SAN DIEGO SUPERCOMPUTER CENTER

The San Diego Supercomputer Center (SDSC) is a leader and pioneer in high-performance and data-intensive computing, providing cyberinfrastructure resources, services and expertise to the national research community, academia and industry. Located on the UC San Diego campus, SDSC supports hundreds of multidisciplinary programs spanning a wide variety of domains, from astrophysics and earth sciences to disease research and drug discovery. Applying themes of "growing a versatile computing ecosystem" and "translating innovation into practice," SDSC continues its leadership with explorations in artificial intelligence, machine learning, cloud and edge computing, distributed high-throughput computing and more.

SDSC INFORMATION
Frank Würthwein, Director

San Diego Supercomputer Center
University of California, San Diego
9500 Gilman Drive MC 0505
La Jolla, CA 92093-0505
Phone: 858-534-5000

info@sdsc.edu
www.sdsc.edu

Cynthia Dillon
Division Director, External Relations
cdillon@ucsd.edu

GROWING A VERSATILE COMPUTING ECOSYSTEM
SDSC Summary Report
July 2021 – June 2023
(PDF version available online at www.sdsc.edu/pub/)

EDITOR:
Cynthia Dillon

CO-EDITOR:
Kimberly Mann Bruch

CONTRIBUTORS:
Kim Bruch, Cynthia Dillon, Jake Drake, Julie Gallardo, Ron Hawkins, Fritz Leader, Susan Rathbun, Bob Sinkovits, Shawn Strande, Ben Tolo, Nicole Wolter

CREATIVE DIRECTOR:
Ben Tolo

PHOTOGRAPHER:
Jake Drake

All financial information is for the fiscal year that ended June 30, 2022. Any opinions, conclusions or recommendations in this publication are those of the author(s) and do not necessarily reflect the views of NSF, other funding organizations, SDSC or UC San Diego. All brand names and product names are trademarks or registered trademarks of their respective holders.

SDSC at Heart

SDSC in Action

SDSC in Academic Research

SDSC in Industry

SDSC in Community

SDSC in Service to Students

SDSC in the Future

DIRECTOR'S MESSAGE

GROWING A VERSATILE COMPUTING ECOSYSTEM

Dear Friends,

As I approach the beginning of my third year as director of the San Diego Supercomputer Center, I understand more than ever how dedicated our organization is to growing a versatile computing ecosystem that can accommodate the needs of the academic research community.

This vision of a "versatile computing ecosystem" is realized through our comprehensive suite of tools and services inclusive of our Expanse supercomputer, Voyager innovative AI system, CloudBank, Prototype National Research Platform, SGCI, SGX3 and more.

With Expanse, SDSC is able to offer composable systems, high-throughput computing, science gateways, interactive computing, containerized computing and cloud bursting. Voyager is designed to facilitate research at scale using artificial intelligence and deep learning, allowing researchers to work with extremely large data sets using standard AI tools, like TensorFlow and PyTorch, or develop their own deep learning models using developer tools and libraries from Habana Labs. CloudBank provides managed services to simplify cloud access for computer science research and education. The Prototype National Research Platform offers a powerful, innovative, nationally distributed system—computing resources, research and education networks, edge computing devices and other instruments—designed as a testbed for science drivers as diverse as the platform itself to expedite science and enable transformative discoveries. Finally, SGCI and SGX3 offer support for the science gateway community of developers.

SDSC's resources are designed to provide the academic research community with the tools needed to conduct research efficiently and effectively, and SDSC's versatile computing ecosystem supports this goal.

Throughout this summary covering highlights from the past two years, you will see examples of how our versatility helps to meet the changing needs of academics, healthcare professionals, industry partners and students. From the day-in, day-out mechanics of running research simulations to storing files, we cover the gamut of needs around high-performance computing, data services, science gateways, cybersecurity, education and training, data science application and more.

I would like to draw your attention to a few feature stories in particular. One is the article about convergence research relative to California's threats of earthquakes, fire and drought (add flooding to the list after recent weather patterns). Another is an update on the Prototype National Research Platform, as well as the National Artificial Intelligence Research Resource Task Force. A third feature – actually a section within the report – is about each of our SDSC divisions. Note the impactful work around cloud computing, cybersecurity and regulatory compliance by our Sherlock Division. The community outreach emphasis of our CICORE Division is also notable. Our service to students and the next generation of data and computer scientists is evident in the information we share here, too.

The summary is packed with information, so my suggestion is to review the table of contents, flip through the pages filled with engaging images and interesting stories and find what appeals to you. I have no doubt that you will come away with your own better understanding of the breadth and scope at which we work every day toward our mission to foster a versatile computing ecosystem.

Happy reading and best wishes,

Frank Würthwein,
SDSC Director

Advanced Technology Lab

In 2022, the Advanced Technology Lab (ATL) at SDSC embarked on a multi-year research program to address the pressing challenges of data movement and management for extreme scale Artificial Intelligence (AI).

The mission of ATL is to identify and evaluate new hardware and software technologies for the future computing and storage systems that will be needed to conduct scientific research at extreme scale. The lab develops research and evaluation programs looking at needs over a three- to five-year time horizon. The lab is funded through gifts from industrial sponsors as well as contracts and grants from federal and other sponsoring agencies.

The computational and storage infrastructure (collectively, "cyberinfrastructure") supporting scientific research is evolving into a globally distributed system ranging from large-scale supercomputer and cloud computing centers to intermediate facilities to edge devices including sensor networks and scientific instruments. More computing capabilities are being incorporated into the edge and the network itself to provide for initial data reduction and analysis. All of this is facilitated by expanding high bandwidth, low-latency research and education networks.

"The increasing use of AI methods on this globally distributed cyberinfrastructure presents challenges with respect to data movement, storage and access at extreme scale. Research and development projects to address these challenges range from QoS and bandwidth guarantees over wide area networks to hardware based 'data accelerators' (DPUs, smart NICs, computational storage, etc.), to 'data preparation for AI,' to software and algorithms that reduce time-to-solution while improving ease of use of the end-to-end systems," said Ron Hawkins, director of Industry Relations at SDSC.

To address the challenges of data management for extreme scale AI, the ATL has begun a three-year program to identify, evaluate and further develop the technologies that will form the global, distributed CI supporting future scientific research. It is expected that many of these technologies will be proposed and potentially deployed in future CI systems funded by the National Science Foundation.

"With the increasing use of artificial intelligence methods to conduct scientific research on a global, distributed computational platform, the management and movement of data at scale becomes a critical enabling factor," said SDSC Director Frank Würthwein."Our goal is to partner with innovators in industry, academia and government to address these challenges."

LEADERSHIP

SDSC EXECUTIVE TEAM

Frank Würthwein
Director, SDSC
Lead, SDSC Distributed High-Throughput Computing
Executive Director, Open Science Grid
Professor, UC San Diego Department of Physics

Ilkay Altintas
SDSC Chief Data Science Officer
Division Director, Cyberinfrastructure and Convergence Research and Education Division (CICORE)

Sandeep Chandra
Division Director, Sherlock

Cynthia Dillon
Division Director, External Relations
Director, Communications

Ron Hawkins
Director, Industry Partnerships (retired June 2022)

Christine Kirkpatrick
Division Director, Research Data Services

Samuel 'Fritz' Leader
Chief Administrative Officer
Division Director, Business Services

Amit Majumdar
Division Director, Data-Enabled Scientific Computing

Shawn Strande
SDSC Deputy Director

Michael Zentner
Division Director, Sustainable Scientific Software

DEPARTMENT LEADS

Chief Information Security Officer
Winston Armstrong

Data Center Operations
Thomas Tate

Desktop Services
Trevor Walker

Facility Services
Sandra Davey

Finance
Julie Gallardo

High-Performance Computing Systems
Christopher Irving

HPC Education, Outreach & Training
Robert Sinkovits
Mary Thomas

Human Resources
Amy Giang-Tran

Multimedia & Social Media
Jake Drake
Benjamin Tolo

News
Kimberly Bruch
Cynthia Dillon

Programs & Events
Susan Rathbun
Cindy Wong

Storage & Compute Services
Brian Balderston

Triton Shared Computing Cluster
Subha Sivagnanam

User Services
Mahidhar Tatineni

Web Services
Michael Dwyer

PI PERSON OF THE YEAR

Ilya Zaslavsky: 2022 Pi Person of the Year

In September 2022, the San Diego Supercomputer Center announced its annual Pi Person of the Year. This is a distinction awarded to an individual at SDSC who consistently demonstrates exemplary research in both science and cyberinfrastructure. The 2022 recipient was Ilya Zaslavsky, director of the Spatial Information Systems Laboratory at SDSC and UC San Diego.

Zaslavsky's research interests focus on distributed information management systems—in particular on spatial and temporal data integration, geographic information systems and spatial data analysis. Among his many accomplishments is the GeoACT (geographically assisted agent-based model for Covid-19 transmission) – a simulation modeling portal designed to help school administrators evaluate risks of COVID-19 transmission in K-12 classrooms and school buses and select appropriate mitigation measures. Zaslavsky, who teaches an upper-division "Spatial Data Science and Applications" course at the Halicioğlu Data Science Institute and a graduate-level "Data Science Approaches to Spatial Analysis" at the School of Global Policy and Strategy at UC San Diego, worked with his undergraduate students to develop the resource, along with physicians at UC San Diego Pediatrics and epidemiologists at the County of San Diego Health and Human Services Agency. An extension of this collaboration was another online tool, the e-Decision Tree, which was used by school administrators and parents from 42 San Diego school districts to generate instructions and manage timing around student and staff exposure, infection, testing and return to the classroom.

Additionally, Zaslavsky has led design and technical development in several cyberinfrastructure projects, including the national-scale Hydrologic Information System, which developed standards, databases and services for integration of hydrologic observations. He was the editor of the initial WaterML specification, and then served, for 10 years, as a co-chair of the Hydrology Domain Working Group of the Open Geospatial Consortium, an international group of experts who developed WaterML 2, the first ever international standard for water data exchange.

Zaslavsky has developed spatial data management infrastructure as part of several large projects in domains ranging from neuroscience (digital brain atlases) and geology to disaster response and regional planning and conservation. He had been an active NSF EarthCube principal investigator (PI), and among many contributions, developed the EarthCube Data Discovery Studio for data search and exploration. His Survey Analysis via Visual Exploration (SuAVE) project focuses on online survey data analysis and has provided value to research projects and communities in many domains. SuAVE was also used for teaching research methods to undergraduate students.

Most recently, Zaslavsky has held PI and senior leadership roles as a part of the NSF Convergence Accelerator program, focusing on Open Knowledge Networks. He actively works on knowledge network applications in biomedical and geospatial domains, contributing also to the SDSC's WIFIRE Commons platform.

"I cannot imagine a more deserving candidate for the Pi Person award," said Ilkay Altintas, founding director of WIFIRE and the division director for CICoRE, SDSC's arm of cyberinfrastructure and convergence research and education. "Ilya has had a lasting impact over three decades, pioneering transformational methods and tools to enable scientific and societal collaborators to effectively apply geospatial data and analysis."

Zaslavsky received his Ph.D. for research on statistical analysis and reasoning models for geographic data from the University of Washington (1995). Previously, he received a Ph.D. equivalent from the Russian Academy of Sciences, Institute of Geography, for his work on urban simulation modeling and metropolitan evolution (1990). Before joining SDSC in 2000, he was a faculty member at Western Michigan University (1995-98). He also worked as GIS staff scientist at San Diego State University (1997-2000) and developed software for online mapping and exploratory social data analysis. Zaslavsky developed one of the first XML-based online mapping systems called Axiomap (Application of XML for Interactive Online Mapping) in 1999.

Over the years, Zaslavsky has earned support from the National Institutes of Health, the National Science Foundation, the U.S. Department of State and other federal agencies, private foundations, Microsoft and the Environmental Systems Research Institute (ESRI).

SDSC BY THE NUMBERS

FISCAL YEAR 2021/2022

ORGANIZATION

$55.4M	ANNUAL REVENUE
$30.6M	GRANT FUNDING
$7.9M	INDUSTRY REVENUE ($300K+ OVER 2020-21)
378	SDSC EMPLOYEES & VOLUNTEERS
54	SPONSORED RESEARCH AWARDS

TRAINING PROGRAMS

48,319	STUDENTS ENROLLED WORLDWIDE IN SDSC-LED ONLINE COURSES
2,152	TRAINING & EVENT PARTICIPANTS
371	STUDENTS ENGAGED IN SDSC PROGRAMS
177	HIGH SCHOOL STUDENTS MENTORED
50	SDSC HOSTED PROGRAMS

SUPPORT FOR UC

698	UC SAN DIEGO USERS: EXPANSE, VOYAGER & COMET
500	UC SAN DIEGO USERS: TRITON SHARED COMPUTING CLUSTER
99	UC SAN DIEGO ACTIVE ALLOCATIONS: EXPANSE, VOYAGER & COMET
30+	PETABYTES OF STORED DATA FOR UC SAN DIEGO & UC

HPC SYSTEMS

200K	x86 CORES ON SDSC HPC SYSTEMS
1,444	ACCELERATORS
528	PUBLICATIONS CITING SDSC/XSEDE RESOURCES
5	COMPUTE CLUSTERS

(Top) SDSC's first supercomputer, the Cray X-MP48 entered production in 1985. (Bottom) Today, SDSC's powerful Expanse system can perform more than five quadrillion calculations per second.

History, Services, Support

SDSC was established as one of the nation's first supercomputer centers under a cooperative agreement by the National Science Foundation in collaboration with UC San Diego and General Atomics Technologies. SDSC first opened its doors on November 14, 1985.

For nearly 40 years, SDSC has cultivated its national reputation as a pioneer and leader in high-performance and data-intensive computing and cyberinfrastructure. Located on the campus of UC San Diego, SDSC provides resources, services and expertise to the local community of researchers, as well as to regional and national partners in academia and industry.

Cyberinfrastructure refers to an accessible, integrated network of computer-based resources and expertise, focused on accelerating scientific inquiry and discovery. With Expanse and Voyager, SDSC's latest supercomputing resources, the center supports hundreds of multidisciplinary programs spanning a wide range of science themes—from earth sciences and biology to astrophysics, bioinformatics and health IT.

Applying the theme "growing a versatile computing ecosystem," SDSC continues its leadership with explorations in artificial intelligence, machine learning, cloud and edge computing, distributed high-throughput computing and more with support from the NSF, NIH, and other agencies and organizations.

SERVICES

SDSC offers a variety of research computing cyberinfrastructure resources, services and expertise. Our Data Center houses a wide range of computational resources that are available to UC San Diego, other University of California campuses and our partner institutions. Among the services SDSC offers:

HIGH-PERFORMANCE COMPUTING - SDSC's HPC experts guide potential users in selecting the right resource, thereby reducing time to solution while taking science to the next level.

DATA SCIENCE SOLUTIONS - SDSC offers complete data science solutions in a breadth of specialties via training, service contracts and joint research collaborations.

CYBERINFRASTRUCTURE SERVICES - SDSC resources for technical research and educational needs include storing public and private data collections, storing sensitive data that is secured to meet regulatory requirements, networking solutions and hosting virtualized platforms, websites and databases.

BUSINESS SERVICES - SDSC's high-tech conference rooms, auditorium, training lab and visualization facilities may be reserved for programs and events.

SUPPORT

SDSC offers in-depth technical support to SDSC service users, including assistance with developing efficient high-performance computing (HPC) applications, prompt service issue identification and resolution, and guidelines that explain how to utilize SDSC resources effectively. Here are some examples:

ACCOUNTS & ALLOCATIONS - SDSC provides a variety of resources and services to the UC/UC San Diego academic research community, national HPC users and industry partners.

RESOURCE DOCUMENTATION – SDSC offers user guides and documentation for its compute and data systems.

TECHNICAL CONSULTING – SDSC's experienced consultants are available to assist users with issues related to SDSC computational resources. We also offer users a 'Helpful Tools' section before submitting questions to "SDSC Consulting."

TRAINING – SDSC supports users of its advanced computing systems, including industry and K-12 and early college users, with education and training programs and events such as conferences, panel discussions, symposia, workshops and two summer institutes.

OVER THE LAST FIVE YEARS

SDSC

HAS RECEIVED

74

NSF AWARDS

TOTALING OVER

98M

DOLLARS

The National Science Foundation and SDSC

The U.S. National Science Foundation (NSF)—an independent federal agency created by Congress in 1950 to promote the progress of science—spurs the nation forward by advancing fundamental research in all fields of science and engineering. The NSF is vital in its support of basic research and the people who do it because the result is knowledge that transforms the future. The NSF also provides facilities, instruments and funding to support researchers' ingenuity and to sustain the U.S. as a global leader in research and innovation.

With a fiscal year 2022 budget of $8.8 billion, NSF funds reached all 50 states through grants to nearly 2,000 colleges, universities and institutions. Each year, NSF receives more than 40,000 competitive proposals and makes about 11,000 new awards. Those awards include support for cooperative research with industry, Arctic and Antarctic research and operations, and U.S. participation in international scientific efforts.

OVER THE LAST FIVE YEARS

SDSC

HAS RECEIVED A COMBINED TOTAL OF

101

GRANT AWARDS

TOTALING OVER

104M

DOLLARS

These totals include the NSF numbers listed on page 8.

Early rendering of the SDSC building before it opened in 1985.

The San Diego Supercomputer Center's relationship with the NSF dates back to 1985 when SDSC was founded with a $170 million grant from NSF's Supercomputer Centers program. SDSC was established as one of the nation's first supercomputer centers under a cooperative agreement among the NSF, UC San Diego and General Atomics (GA) Technologies.

Since then the NSF and SDSC have worked together on numerous grants for projects that honor the NSF's goals of promoting science and advancing fundamental research across all fields of science and engineering.

Today, advanced digital technology has become ubiquitous in the daily lives of researchers in science, engineering, social science and the humanities around the world. Resources such as supercomputers, collections of data and emerging tools are vital to the livelihoods of users whose work improves lives.

NSF, with its continued generous support of projects such as the Extreme Science and Engineering Discovery Environment (XSEDE) and the new $52 million Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, provides users with seamless integration to NSF's high-performance computing, data resources and digital services, as well as its cyberinfrastructure ecosystem.

From September 2019 to June 2022, SDSC's Chaitan Baru was on assignment at the NSF as Senior Science Advisor in the Office of Integrative Activities, Office of the Director. He provided technical leadership for the Open Knowledge Network of the NSF Accelerator. Between August 2014 and 2019, he served as the first (and only) Senior Advisor for Data Science in the Computer and Information Science and Engineering Directorate (CISE). He co-chaired the NSF Harnessing the Data Revolution Big Idea, played a leadership role in the NSF BIGDATA program and advised the NSF Big Data Regional Innovations Hubs and TRIPODS programs. He also helped to establish the partnership between the NSF BIGDATA program and the public cloud providers–AWS, Google, Microsoft and IBM in 2017. Baru was a distinguished scientist at SDSC but now serves as Senior Advisor in the new NSF Directorate for Technology Innovation and Partnerships (TIP).

SDSC AT HEART

The San Diego Supercomputer Center is a leader in high-performance and data-intensive computing and cyberinfrastructure. SDSC provides resources, services and expertise to the national research community—from academia to industry. With its robust and dependable array of resources, SDSC serves hundreds of multidisciplinary programs spanning a wide variety of domains—from earth sciences and astrophysics to bioinformatics and health IT. In this section, we invite you to learn more about our computing and cyberinsfrastructure resources and how they are used to impact society--from addressing California's triple threat of fire, drought and flooding, and earthquakes to improving weather forecasts and democratizing cyberinfrastructure.

Expanse Supercomputer Spotlights Role of Unstable Cells in Response to TB Infection

According to a 2021 World Health Organization report, the global COVID-19 pandemic caused an increase in tuberculosis (TB) deaths – 1.5 million in 2020 versus 1.4 million in 2019 – due to a lack of efficient diagnosis and treatment. Since TB has been the number one cause of death from infectious disease in the world for centuries, University of Michigan (UM) Medical School Professor Denise Kirschner and colleagues have been working to better understand the disease by using supercomputers like Expanse at SDSC.

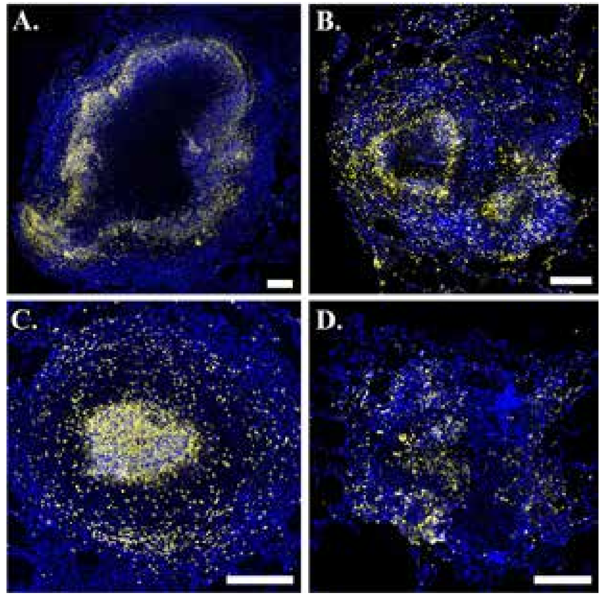
“Even during this COVID-19 pandemic, TB remains the leading cause of death by infectious disease around the world,” said Kirschner, a mathematical biologist, whose latest findings were published in the Frontiers in Immunology journal. “Until our recent study, a particular group of cells called neutrophils have been somewhat of a black box, but our latest work has allowed us to show the potential role of these cells in the immune response to TB infection.”

Kirschner explained that neutrophils are unstable cells with short life spans when grown in the lab, but predictive simulations using the Expanse supercomputer allowed her team to use images of the cells inside of a TB granuloma to create high-resolution models illustrating how these mysterious cells react when infected with TB.

“The immune system is made of up of many cells that help respond to an invasion of microbial pathogens and advancements in experimental science gave us imaging of these cells in the granuloma so that we could use that information to inform our model,” she said. “Our new simulations helped us illustrate that there must exist a dual role of neutrophils, which explains why they have been so difficult for us to study and predict their behavior.”

With these new simulations, Kirschner said that the research team now has both mechanisms and information regarding the role of neutrophils in the immune response to infection during tuberculosis. Specifically, their new computational models showed how the host process contributes to immune and drug dynamics—ranging from the molecular level to the entire host.

This research was supported by grants from the National Institutes of Health, Bill and Melinda Gates Foundation, and the Wellcome Trust. Computations on Expanse were allocated by XSEDE (TG-MCB140228). Data derived from The University of Pittsburgh non-human primate tuberculosis lab of JoAnne Flynn and Joshua Matilla made this work possible.



The hallmark of TB infection is the formation of spherical structures in the lungs. Theses masses of infected tissue, called granulomas, start to form within the first two to four weeks after infection. The yellow portions of these illustrations show the role of neutrophils (unstable cells) at various stages of TB infection. Image source: Frontiers in Immunology

Frontiers in Immunology: Neutrophil Dynamics Affect Mycobacterium tuberculosis Granuloma Outcomes and Dissemination

www.frontiersin.org/articles/10.3389/fimmu.2021.712457/full

Why it's important

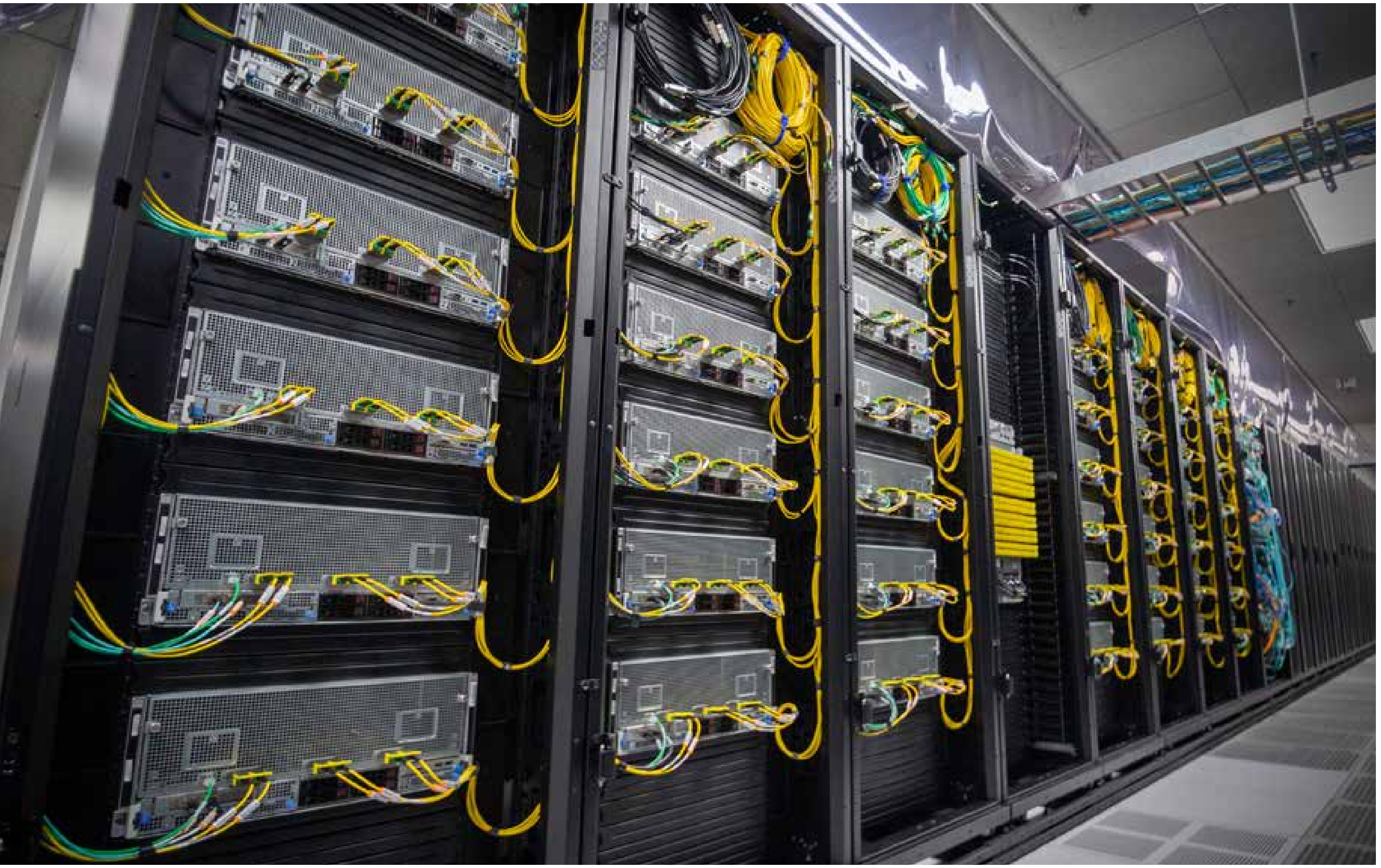
Any scientific progress made toward understanding how to improve immunity to TB infection can greatly impact the world.

According to researchers, infection results from inhalation of the bacteria Mycobacterium tuberculosis (Mtb). While TB is treatable, it requires multiple antibiotics taken for more than six months, and the emergence of drug resistant Mtb has strained the current arsenal of effective TB drugs. The situation is critical considering there are nine million new cases of active TB every year and over one million annual deaths.

The pathological hallmarks of TB are lung granulomas, which are dense spherical collections of immune cells that serve to protect the host by isolating bacteria; however, this also provides a niche for bacterial survival and growth. Because granulomas are dense heterogeneous structures, they also present a physical barrier for antibiotic diffusion, slowing drug penetration. These difficulties contribute to the challenge of devising new and more effective treatment strategies for TB: getting the right drugs at the right concentration to the right location to kill the appropriate bacterial subpopulation.

Additionally, while there is a vaccine for TB, known as BCG, it is not used in the U.S. because of its variable efficacy and nullification of available TB tests; a highly effective vaccine still remains elusive though there are currently 10 candidate vaccines in trial.

“Our computational approach and results from our recent study have helped narrow this concerning ‘vaccine design space’—getting us closer to a globally effective vaccine,” Kirschner said. “Developing and testing these models and using them to make critical predictions about TB treatment and prevention, is a decade-long project and we are hopeful that this current research gets us that much closer to a vaccine that can be used throughout the world.”



Experimental Voyager Enters Testbed Phase

Voyager, the high-performance/high-efficiency, experimental compute resource installed early in 2022 at SDSC and sanctioned for production by the National Science Foundation (NSF), moved into its operational testbed phase a few months later in 2022.

Envisioned as a system to facilitate exploration of new architectures in support of artificial intelligence (AI) in research and engineering, Voyager is a major departure from former NSF systems that have been focused on delivering computing resources in support of traditional applications and programming models. Voyager instead emphasizes deep engagement with the AI research community and features specialized hardware and software, close collaboration with applications teams and the opportunity to share these experiences within the community.

According to Voyager Principal Investigator Amit Majumdar, the NSF-supported Voyager project is structured in two phases: 1) a three-year testbed phase and 2) a two-year allocations phase.

“The testbed phase is centered around deep user engagement, whereby select research groups will provide information to help evaluate Voyager’s innovative deep learning (DL) hardware, software, libraries and machine learning (ML) application porting and performance,” said Majumdar.

The testbed phase is guided by an External Advisory Board that helps recruit research groups. During the first years, the project will offer semiannual workshops and user forums to share lessons learned and to bring researchers together. These approaches will help to develop a knowledge base, best-use cases for future users and allocation policies. The allocations phase will follow via an NSF-approved process, which will be informed by the lessons learned from the testbed phase, regular and advanced user support, semiannual workshops and industry engagement for similar technology evaluation.

“The National Science Foundation is delighted to see the Voyager system move into its operational testbed phase,” said Manish Parashar, NSF director, Office of Advanced Cyberinfrastructure (OAC). “Artificial Intelligence research is playing an increasingly important role across all areas of science, engineering research and education. Voyager, with its specialized hardware and software capabilities and deep engagements, can be a tremendous resource for the community, providing new research opportunities and driving innovation.”

According to Majumdar, Voyager’s architecture features hardware and software innovations that will lead to performance gains and ease porting and model development in AI. Voyager comprises 42 Supermicro X12 Gaudi® AI Training Systems with 336 Habana Gaudi processors—designed for scaling large supercomputer training applications—and 16 Habana first-gen inference processors to power AI inference models. Voyager’s networking is designed to support very large AI models. Each Gaudi processor has 10 integrated ports of RoCE (RDMA over converged Ethernet), with the 42 training systems connected by six 400Gpbs connections into a large Arista non-blocking switch.

“Porting of user applications has been relatively straightforward,” said Majumdar, adding that several applications are now running on Gaudi and inference processors. “So far our experience is that codes need minimal changes.”

With support from Habana developers, in collaboration with SDSC and researchers, the porting of user applications is providing a basis for training materials, including a three-hour session that has been recorded for other users.

“One team has been running applications via Jupyter Notebook on Gaudi, and users can run with familiar TensorFlow and PyTorch frameworks,” said Majumdar, noting that the system was designed to support exploration in multiple dimensions (Gaudi and inference processors, RoCE interconnect, 400 GbE switch, Kubernetes, Ceph, cnvrg.io, Slurm and more).

According to SDSC’s Deputy Director Shawn Strande, strong collaboration with technology partners at Supermicro and Habana allowed SDSC to bring this innovative architecture to the community.

“Considering the amount of innovation, the acquisition, deployment and installation has been remarkably smooth (even in the midst of Covid-19 conditions), with issues resolved jointly with Supermicro and Habana experts,” said Strande, who also serves as the project manager for Voyager.

SDSC experts reported that in most cases, measured performance has been better than projected, due in large part to software improvements by Habana and the collaboration.

“Supermicro is excited to continue supporting SDSC’s multi-year Voyager AI project as it enters the critical testbed phase of production operations,” said Ray Pang, vice president of technology enablement at Supermicro. “Supermicro’s ability to create complex AI solutions encompassing networking, compute, storage and AI, demonstrates how Supermicro’s AI and HPC solutions are ideal for science, medical and academic research.”

Fire, Water and Earthquakes: California's Triple Threat

WIFIRE LAB, WIFIRE COMMONS AND BURNPRO3D: A HOT TRIO

A critical component of SDSC's Cyberinfrastructure and Convergence Research and Education (CICORE) Division is the WIFIRE Lab–a consortium of UC San Diego organizations and a number of partnerships including university collaborators, industry partners, fire departments, utilities, the California Governor's Office of Emergency Services (Cal OES) and the California Public Utilities Commission.

To meet growing needs in hazards monitoring and response, the WIFIRE Lab, led by SDSC's Chief Data Science Officer Ilkay Altintas, is an all-hazards knowledge cyberinfrastructure (CI). Currently, it is the only integrated infrastructure of its kind that can be a neutral data resource/partner to any proposed activity. The WIFIRE Lab adds value to the raw data and prepares the best data in real-time for any monitoring and modeling effort—as well as its dynamic data-driven models—for research and operational use.

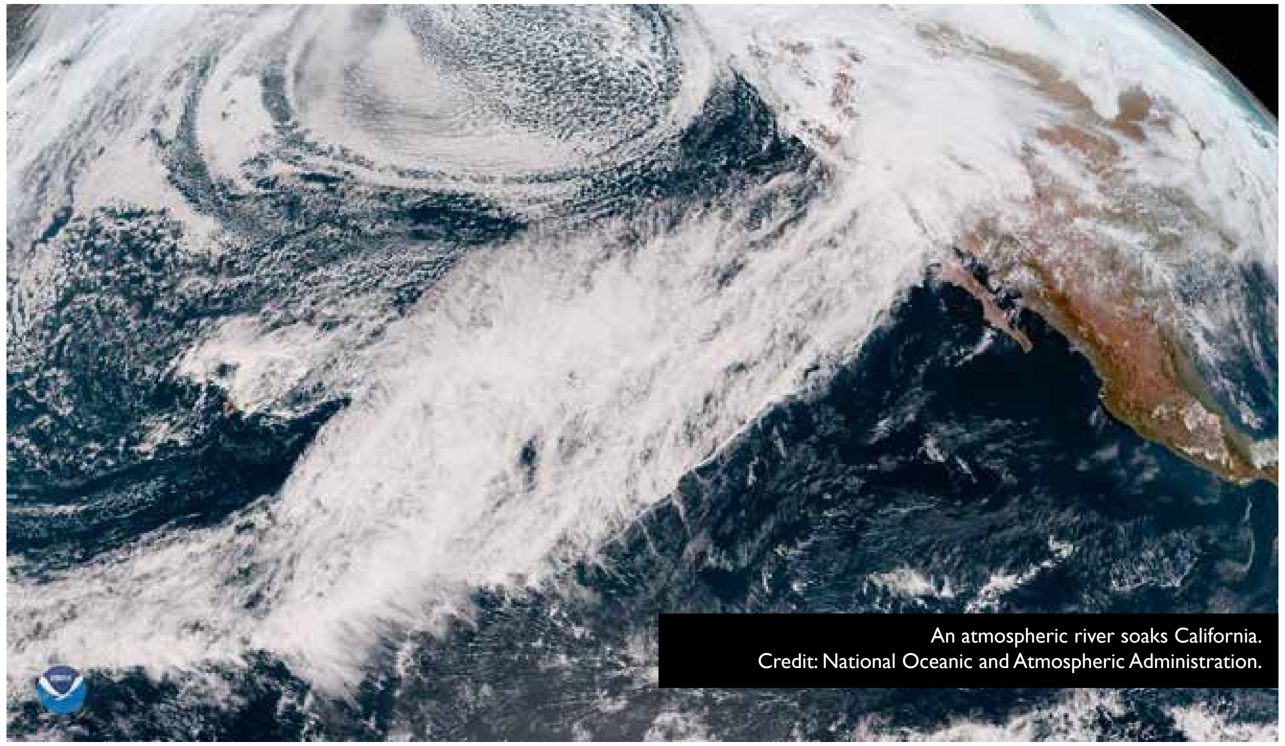
While WIFIRE Lab focuses on CI, the WIFIRE Commons enables AI-driven societal and scientific wildland fire applications through data and model sharing. The primary objective of the project is to create a convergence environment to accelerate wildland fire science and its proactive application to operational use for mitigation, planning, response and recovery through artificial intelligence (AI) innovations.

To achieve convergence between AI and fire science communities, WIFIRE Commons develops an intelligent and integrated infrastructure to catalog, curate, exchange, analyze, optimize and communicate big data and models at scale.

New to CICORE is BurnPro3D—a decision support platform to help the fire response and mitigation community understand risks and tradeoffs quickly and accurately so that they can more effectively manage wildfires or conduct controlled burns. The innovative platform that brings science, data and technology together in dynamic fashion is funded through the NSF's Convergence Accelerator program.

Leveraging the WIFIRE Commons data-sharing and AI framework, the BurnPro3D platform uses next-generation fire science in prescribed burns for preemptive vegetation treatment at an unprecedented scale. Altintas is the project's principal investigator.

"BurnPro3D harnesses the data and AI capabilities in WIFIRE Commons for optimization of fire mitigation efforts. The NSF convergence accelerator program is all about innovation for societal impact. We have been developing key infrastructure and partnerships in this area for the last eight years, and more recently working with our BurnPro3D collaborators to include next-generation fire science and AI in various aspects of the project," said Altintas.



An atmospheric river soaks California.
Credit: National Oceanic and Atmospheric Administration.

COMET SUPPORTS IMPROVING WEATHER AND HYDROLOGICAL FORECASTS

During summer 2021, SDSC's petascale Comet supercomputer completed its formal service as an NSF resource and transitioned to exclusive use by Scripps' Center for Western Weather and Water Extremes (CW3E). The transition enabled CW3E researchers to leverage Comet's computing capabilities—nearly 3 quadrillion operations per second—to improve weather and hydrological forecasts with the goal of enhancing the decision-making process associated with reservoir management over California. Anticipated results of the transition include increased water supply and reduced flood risk over the region.

In early 2022, the CW3E team conducted a "Near Real Time" 200-member WRF Ensemble run on Comet in support of its Atmospheric River Recon campaign. The ensemble was at a scale much greater than most groups run—it used about 1,200 nodes per night and, given a daily 7:00 a.m. deadline, completed in under 10 hours. The team was successful in getting nearly all members out per night.

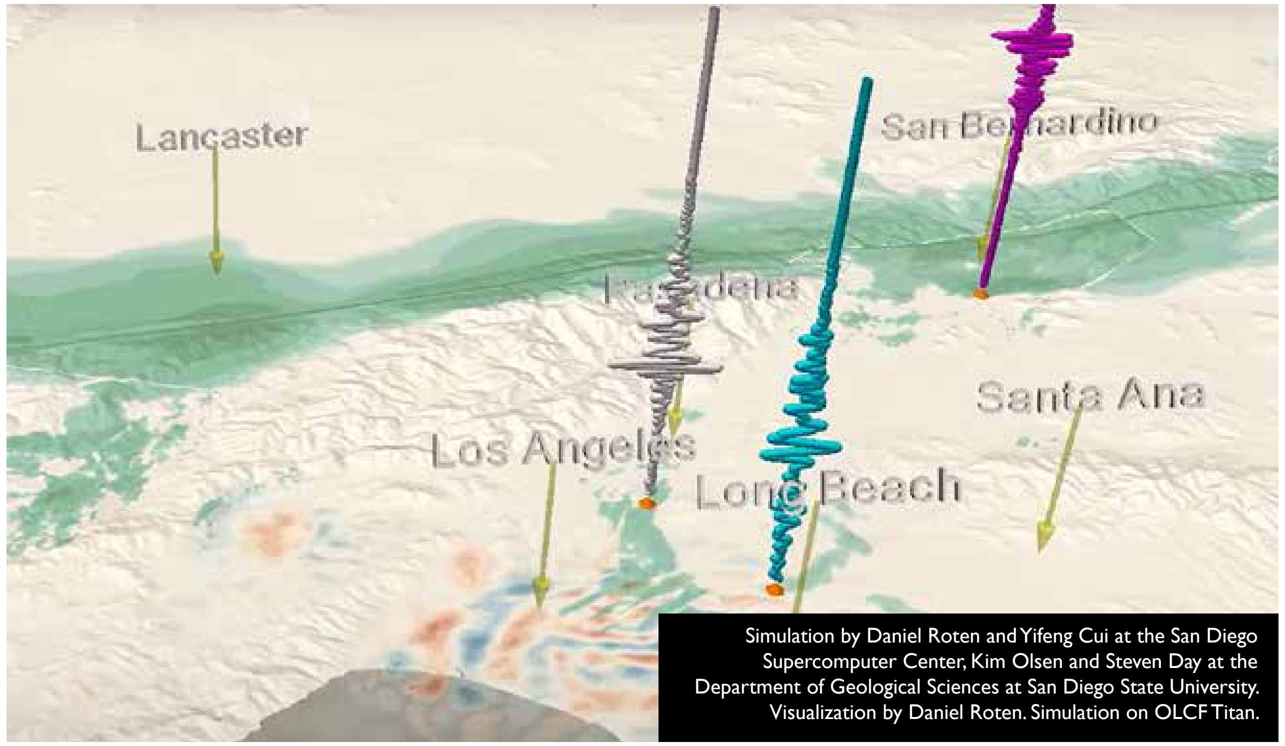
"We are still working on an analysis of the benefit of such a large ensemble when predicting extreme events, but early analysis shows promise," said Patrick Mulrooney, CW3E's domain science programmer.

According to Luca Delle Monache, academic program manager of climate, atmospheric science and physical oceanography at Scripps, the exclusive access to Comet allows CW3E to advance several computational projects that would not be possible otherwise.

"The ability to perform computational experiments, that just one year ago we could not even think of, and to strengthen our collaboration with key partners as the National Weather Service (NWS) is an exciting aspect of this agreement," Dell Monache said.

CW3E's computational projects span a broad scope from basic to applied research. The projects include: trying to answer fundamental questions on the formation and evolution of atmospheric rivers, and the interactions between the atmosphere and the ocean; studying orographic precipitation, generated when the moist air mass associated with an atmospheric river is lifted over mountains and condenses into rain or snow; developing high-resolution, sub-seasonal to seasonal predictions; testing and improving operational systems run from the NWS; and expanding CW3E near real-time operational capabilities for weather and hydrology with high resolution deterministic prediction and an ensemble with an unprecedented large number of members.

"We will also develop and test new machine learning algorithms and dynamical models leveraging Comet's several graphics processing unit (GPU) nodes," said Delle Monache. "Moreover, we will develop and test new data assimilation schemes."



Simulation by Daniel Roten and Yifeng Cui at the San Diego Supercomputer Center, Kim Olsen and Steven Day at the Department of Geological Sciences at San Diego State University. Visualization by Daniel Roten. Simulation on OLCF Titan.

SOUTHERN CALIFORNIA EARTHQUAKE CENTER WELCOMES SDSC

For more than three decades, the Southern California Earthquake Center (SCEC) has been a consortium of universities and scientific institutions—one of the largest research collaborations in geoscience. It performs fundamental research in earthquake processes using southern California as its principal laboratory.

While SDSC has been a close collaborator of the SCEC community for 20 years—particularly in areas of large-scale earthquake simulation and application development—the SCEC Board of Directors recently designated SDSC as a core institution.

"We couldn't be more pleased by the opportunities presented to our team with this partnership as a core institution with SCEC," said SDSC Director Frank Würthwein. "We are honored to play a vital role in this important work."

SDSC has helped SCEC develop some of the most efficient and scalable earthquake simulation software on both CPU- and GPU-based architectures. For instance, an open source code co-developed by SDSC was used in a Gordon Bell prize-winning simulation in 2017, which is presented each year by the Association of Computing Machinery at the annual Supercomputing Conference. Another version of this code, providing equivalent results, is used as a workhorse with GPU acceleration to calculate ground motions from many single-site ruptures for CyberShake, a SCEC computational platform that generates the physics-based California statewide seismic hazard map.

"While we have a track record of joint efforts for end-to-end earthquake simulations in the past, our new role will strengthen research computing and data science across the center's activities, while incorporating additional resources for training and guidance on both hardware and software, to support a globally renowned leader in earthquake science research," said Yifeng Cui, a computational scientist at SDSC who has been appointed as the institutional representative for the SCEC Board of Directors.

According to Yehuda Ben-Zion, SCEC director and earth sciences professor at the University of Southern California, SCEC and SDSC have had a long and successful partnership. "We very much look forward to collaborating on expanded projects from cyberinfrastructure development, machine learning and AI, to community engagement that benefit from our complementary strengths," he said.

Additional activities will involve collaboration between SCEC and the Cyberinfrastructure and Convergence Research and Education (CICORE) Division at SDSC, led by Ilkay Altintas.

Democratizing Cyberinfrastructure ACCESS for Researchers

Science and engineering research and education depend on a complex and distributed ecosystem of cyberinfrastructure (CI). This ecosystem is made up of research labs, campuses and national resources. In an effort to support this evolving and expanding environment, the National Science Foundation (NSF) has initiated its new $52 million Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program.

As prescribed in NSF's Transforming Science through Cyberinfrastructure Blueprint, this evolving system needs a governance model that is forward-looking and guarantees highly reliable systems and services that the STEM community depends upon. At the same time, the model requires flexibility and structures to adapt as the ecosystem and community continues to evolve. ACCESS will help achieve this through a set of distinct service tracks and an ACCESS Coordination Office (ACO) to assist the multiple service tracks with governance, communications and evaluation.

SDSC's role in the project is to apply its "resources and expertise with long-time collaborators—the National Center for Supercomputing Applications (NCSA) and the Center for Education Integrating Science, Mathematics and Computing (CEISMC) at the Georgia Institute of Technology—to operate Open CI Ecosystem to Advance Scientific Discovery (Open CI), specifically the ACO.

The ACO team includes SDSC's Deputy Director Shawn Strande, ACO Principal Investigator (PI) John Towns (NCSA), Co-PI Lizanne DeStefano (Georgia Tech) and other team members.

"OpenCI will foster an environment for the ACCESS program where shared governance and horizontal leadership create an inclusive and vibrant CI ecosystem in which service track PIs work with common purpose through well-defined decision-making processes, transparency in communication and a singular focus on enabling science," said Strande, co-PI for ACCESS.

As a sub awardee, SDSC helps manage the ACO over a five-year period, which began July 1, 2022 and ends June 30, 2027, with funding of $1.25 million. SDSC also plays a vital role in the development and facilitation of the ACCESS External Advisory Board, development and publication of science communication pieces and external evaluation of the ACO itself.

SDSC is also a partner on a second award led by Amy Schuele at NCSA, the COre National Ecosystem for CyberinfrasTructure (CONECT), which provides SDSC with $1.4 million over its five-year duration. This project encompasses operations and integration services for one of the service tracks.

"SDSC contributes in two main areas," said Robert Sinkovits, CONECT PI and SDSC's director of education and training. "Cybersecurity expert Scott Sakai plays a key role in risk management, federated intelligence sharing and CONECT authentication strategies. Tom Hutton, a highly experienced networking engineer who is widely known for his work on SCINet, the high-performance network built each year for the SC conference, supports CONECT's efforts in application metrics along with networking resources and tools. Tom will also be responsible for mentoring a junior network engineer."

Other partners in CONECT are Florida International University, Indiana University, Pittsburgh Supercomputing Center and the University of Chicago.

SDSC's role extends further to a third award for a $10 million project led by the University at Buffalo (UB) to provide monitoring and measurement services using their software program called XD Metrics on Demand (XDMoD), which is used widely in academia, industry and government agencies to manage high-performance computing infrastructure.

"Our role in this proposal is to help integrate public cloud monitoring data from the NSF CloudBank project into XDMoD's analytical tools, which will help provide a more comprehensive view of the nation's cyberinfrastructure ecosystem and better facilitate planning," said Shava Smallen, computational and data research specialist at SDSC and Co-PI of the NSF CloudBank project.

Thomas Furlani, PI for the project and chief information officer at Roswell Park Comprehensive Cancer Center, as well as research associate professor of biomedical informatics in the Jacobs School of Medicine and Biomedical Sciences at the University at Buffalo, coordinates the work of partners at SDSC, Case Western Reserve University, Indiana University, the University of Texas at Austin and Tufts University.

ACCESS is the next-generation system to the NSF's Extreme Science and Engineering Discovery Environment (XSEDE), a single virtual system established in 2011, which transitioned to ACCESS in early September 2022. ACCESS connects U.S. scientists to supercomputer resources and services nationwide, transforming scientific exploration by putting increasingly powerful machines at the disposal of new communities of investigators.

Additional National Impacts

PNRP INSTALLED AT UC SAN DIEGO

The Protype National Research Platform (PNRP) has been recently installed in eight locations nationwide, including SDSC at UC San Diego. Additionally its acquisition review has been completed, moving it into formal operations as a testbed for exploring a wide range of hardware and new approaches for moving data over high-performance content delivery networks.

For this first-of-its-kind resource, the NSF awarded SDSC $5 million in hardware and $7.25 million for operations over five years. The award supports hardware and deployment across three facilities: on the East Coast at the Massachusetts Green High Performance Computing Center in Mount Holyoke, MA; in the Midwest at the University of Nebraska–Lincoln and on the West Coast at SDSC—as well as five data caches in the Internet2 network backbone. Recently installed in all eight locations nationwide and its acquisition review completed, PNRP has entered formal operations as a testbed for exploring a wide range of hardware and new approaches for moving data over high-performance content delivery networks.

The system design includes 8 DGX with 8 A100 80 gigabytes, the capability to connect four of the DGX at 8xGen4 PCIe max bandwidth, 32 Xilinx U55C connected with two Ethernet NICs to a switch and much more. Racks for the PNRP were installed for SDSC at various locations on the UC San Diego campus during summer 2022. With the system installed and operating, SDSC hosted a workshop in February 2023, establishing an AI/ML community platform.

"PNRP provides the foundation for the NRP community, a platform built by the community for the community," said SDSC Director Frank Würthwein.

In 2021, SDSC responded to the NSF's call for a cyberinfrastructure ecosystem that meets the needs of today's data-intensive science with PNRP. The innovative testbed system includes computing resources, research and education networks, edge computing devices and other instruments that will expedite science and enable transformative discoveries across diverse science themes.

NAIRR TASK FORCE UPDATE

In June 2021, the Biden Administration announced the National Artificial Intelligence Research Resource (NAIRR) Task Force—a group of 12 individuals from academia, government and industry, including former SDSC Director and Distinguished Professor of Physics at UC San Diego Michael Norman. The task force's charge was to propose a vision and implementation for NAIRR that supports access to federal data by researchers in order to keep the U.S. at the forefront of emerging technology.

Since June 2021, the NAIRR Task Force convened to make progress toward the envisioned goals. The Task Force submitted an interim report to the President and Congress in spring 2022, outlining a general vision for how NAIRR could be structured, designed, operated and governed to meet the needs of America's research community. In the report, the task force presented an approach to establishing the NAIRR that builds on existing and future federal investments; designs in protections for privacy, civil rights and civil liberties; and promotes diversity and equitable access.

Concurrently with the publication of the interim report, the task force issued a Request for Information (RFI) to solicit public feedback on the task force's preliminary findings and recommendations outlined in the report, and particularly, on how the recommendations could be successfully implemented.

Public responses to the RFI were accepted through June 2022. The task force integrated this public feedback with other inputs and further deliberated with a top-of-mind goal to democratize access to resources and tools that fuel AI research and development, expanding the ability of academia and government to explore innovative ideas for advancing AI, grounded in ethical principles, in a range of scientific fields and disciplines. The task force issued its final report in January 2023. The proposed vision outlined a set of shared computing and data infrastructures to provide AI researchers and students with compute resources and high-quality data, along with appropriate educational tools and user support.

Our Divisions & Centers of Excellence

FROM FINANCE AND HR TO FACILITIES AND DATA SYSTEMS, SDSC BUSINESS OFFICE GETS IT DONE

If the San Diego Supercomputer Center (SDSC) has a superhero it just might be the Business Services Division. As the center's business office, the division provides the administrative support for the entire center through four main areas: finance, human resources (HR), facilities and data systems.

"We provide all aspects of HR support—guiding employees from their first day to their last. We provide financial support and controls in the form of center budgeting, post-award support to contracts and grants, invoicing and management of service agreements, and management of self-supporting activities, to name a few items," said Fritz Leader, division director. "We support center staff in all purchasing and expense reconciliation. We provide pre-award support for proposals and contracts in development. We care for the building facilities, coordinating maintenance, moves and allocations of space. We also develop business applications to fill gaps not provided by campus services."

The list of services the business office provides is long, but Leader describes the division's impact by using grants as an example.

"Every large grant or contract that SDSC receives—and is known for—has started as a proposal the business office has managed and submitted. For each of those grants to be successful, the business office provides the fiscal management needed to meet sponsor goals. Every research staff member who works on a grant has been hired and onboarded by our HR group and works in building space that we manage. All of this activity is tracked and managed in both campus systems and applications we develop. In short, we have broad impact across all of the center's activities," he said.

One of the heroes within the division is Sandra Davey, who manages Facility Services. Throughout the course of the COVID-19 pandemic, Davey worked on site every day of every work week—sometimes on weekends, too.

"I'm honored to be a long-term employee of SDSC. Many times, I may be the first person a new employee or visitor will meet; it is a privilege to be the first face they see representing SDSC," said Davey. "I am also fortunate to interact with a large group of essential workers at SDSC and UC San Diego. Together, we provide the 'get it done' services that are required in the day-to-day operation of SDSC."

Leader says that what he appreciates the most about the division he directs is the people, whom he describes as highly competent with a cohesive sense of team and a genuine care for each other.

"It's comforting knowing that we are all available to support each other on any issue—and we have had a landslide of them over the past two years. We have been through a lot of overlapping UC San Diego systems transitions. As a group we actively share together what we learn about emergent challenges and learn from what everyone is experiencing. But despite challenges, I also know that we can continue to laugh together and enjoy each other's company," said Leader.

Designing California's Future" was a collaboration between SDSC's CICORE Division and the UC San Diego Design Lab where 23 teams (88 students) participated in a Design-a-Thon to create proactive solutions to end destructive wildfires.

BUILDING CYBERINFRASTRUCTURE SYSTEMS FOR IMPACT

The Cyberinfrastructure and Convergence Research and Education (CICORE) Division at SDSC combines broad expertise in cyberinfrastructure (CI) with deep, domain-specific expertise in artificial intelligence (AI)-enabled science to build use-inspired solutions to grand societal challenges at scale with partners in research communities, practical communities and industry. Projects are led by the division's world-class experts in data science, computing, workflow management, GIS, knowledge networks and AI, as well as thematic areas such as earthquakes, wildfires and genomics.

"Our culture of problem solving and real-world impact, as well as the experiential education opportunities we provide, make me very proud of our team's work," said CICORE Division Director Ilkay Altintas, who is also a Founding Fellow of the Halicioğlu Data Science Institute and the Chief Data Science Officer at SDSC.

CICORE initiatives address a wide range of issues from food security to cybersecurity. What the work has in common is the need for cyberinfrastructure and AI and machine learning (ML) approaches that are designed to deal with information from multiple sources, complex data and many different users with varying needs. The primary groups within CICORE, along with their team leads are as follows:

- Advanced Cyberinfrastructure Development Lab - Vishu Nandigam
- Advanced Query Processing Laboratory - Amarnath Gupta
- BlockLAB - James Short
- Center for Applied Internet Data Analysis - Kimberly Claffy
- Convergence Research (CORE) Institute - Zaira Razu
- High-performance GeoComputing Lab and Intel Parallel Computing Center - Yifeng Cui
- Spatial Information Systems Laboratory - Ilya Zaslavsky
- Structural Bioinformatics Laboratory - Peter Rose
- Workflows for Data Science (WorDS) Center of Excellence and WIFIRE Lab - Ilkay Altintas

One example of an impactful CICORE project involved the creation of "COVID Decision Trees," which were developed in a long-term partnership between CICORE researchers and multiple UC San Diego collaborators and governmental agencies. A project led by CICORE Researcher Ilya Zaslavsky, along with several UC San Diego undergraduate data science students, who developed an agent-based simulation system to assist in COVID-safe school re-openings within San Diego County. Zaslavsky and the team created a spatially explicit, agent-based modeling of COVID-19 transmission at schools and on school busses – allowing individual sites and districts to test their plans and match them with their specific spaces, resources and population.

Another example of CICIORE researchers collaborating for real-world impact is the work of the WIFIRE Lab to provide a data infrastructure and solutions for wildfire response and mitigation efforts, becoming a management layer from the data to knowledge generation and modeling efforts for government agencies and utilities alike.

"In our age of complex societal-scale problems, there exists a growing need for university researchers to participate in multi-sector and cross-disciplinary partnerships focused on impact," said CICORE Director of Strategic Partnerships Melissa Floca. "Convergence research in data science and computing requires the use-inspired, team science approach that our division brings to our work with the fire management community."

A new CICORE Initiative that exemplifies the work of the division is the recent launch of the Convergence Research (CORE) Institute. The CORE Institute is a yearlong training program that is designed to catalyze an impact network of researchers, practitioners, and industry and public policy professionals committed to collaboratively engaging in convergence research.

"The long-term success of CICORE is built around three pillars – our collaborative culture and partnerships, our expertise in cyberinfrastructure and data science, and our cross-disciplinary approach to solving problems. In the CORE Institute, we aim to share these with our fellows," said Altintas.

The institute's theme for 2023 is Tackling Climate-Induced Challenges with AI. Fellows will transfer ideas and technologies to practice and design AI solutions for climate change adaptation, resilience and/or mitigation.

"CICORE's ability to create impact is a product of the focus from project inception on not just innovation but also on building intentional pathways to scale and sustainability," said Chaitan Baru, who until recently was a distinguished scientist at SDSC and is now Senior Advisor in the new NSF Directorate for Technology, Innovation and Partnerships (TIP).

DESC team members and others pictured at a division event, March 2021. Photo courtesy of Amit Majumdar.

DIVING INTO DATA-ENABLED SCIENTIFIC COMPUTING AT SDSC

In addition to SDSC's NSF-funded supercomputers, grants acquired by the Data-Enabled Scientific Computing (DESC) Division's principal investigators (PIs) and Co-PIs span a range of areas—high-performance computing (HPC), molecular dynamics, quantum chemistry, genomics, cosmology, neuroscience, data provenance, parallel math libraries, commercial cloud access, CI software, HPC network, science gateways and education. This diversity in grant domains parallels the breadth of the DESC Division, which is one of seven at SDSC.

Headed by Amit Majumdar, DESC is organized into multiple groups which have specific expertise to lead HPC and computational sciences innovation and to serve and collaborate with thousands of SDSC's national, University of California-wide and UC San Diego researchers, as well as industry partners.

"We provide full support to the user community, and provide training to thousands of users on various topics that allow them to use these machines effectively to do science and education, including classroom teaching," said Majumdar. "We also develop and improve performance of user applications covering the complete set of science domains such that the codes make optimal use of our compute resources in terms of scalability of compute algorithms with massive parallelism, data movement and data storage. The science gateways, used by thousands of researchers, use these supercomputers for computing."

In recent years SDSC supercomputers have enabled the "long tail of science" and "computing without boundaries" for about 100,000 users for research and education. "We strongly contribute toward democratization of access to large-scale computing, including the commercial cloud, by researchers and educators nationally," said Shava Smallen who leads the Cloud Software Development Group.

According to Majumdar, DESC has about 30 staff members and about two-thirds of them have master's degrees or doctoral degrees in computational/computer science, domain science areas of biochemistry, bioinformatics, data science, physics and many engineering disciplines. DESC researchers have published their results in journals such as Nature and Science; they present at reputable national and international conferences, often winning best paper recognitions and awards such as the Gordon Bell Prize.

"Many of the staff members wear a dual hat, where, in addition to working on supercomputers or projects funded by NSF, UC San Diego, UC or industry, they are PIs and Co-PIs on research grants funded by NSF, NIH and other funding agencies," said Subhashini Sivagnanam, who leads the Cyberinfrastructure Services and Solutions Group.

According to Andrea Zonca, who leads the Scientific Computing Applications Group, the DESC division functions in such a way that allows it to combine the knowledge, expertise and experience of DESC members from different groups. "This enables us to provide optimal solutions to SDSC's large user communities, be it by developing and operating supercomputers and clusters or via funded research projects led by PIs and Co-PIs from DESC," said Zonca.

DESC also utilizes expertise and resources from other SDSC divisions and groups such as Research Data Services (RDS), Cyberinfrastructure and Convergence Research and Education (CICORE), Sustainable Scientific Software (S3), High Throughput Computing (HTC), Sherlock Cloud Solutions and Services (SCSS), Business Services and External Relations.

"The depth and breadth of expertise and knowledge of DESC staff members, and the research carried out by many of them, allow us to serve the broad community of researchers in academia and industry," said Christopher Irving, leader of the High-Performance Computing Systems Group. "All of our groups serve them by operating and supporting the supercomputers and by providing CI solutions for research and education."

Supercomputers and clusters, which DESC designs, builds and operates, include the Expanse supercomputer; Comet, previously funded by the NSF and currently funded by the Center for Western Weather and Water Extremes (CW3E) at Scripps Institution of Oceanography; the Triton Shared Compute Cluster (TSCC), an agile, medium-scale, high-performance and efficient computing cluster primarily for campus researchers and students to explore and begin innovative research and the Popeye machine, operated at SDSC on behalf of the Flatiron Institute of the Simons Foundation. There are two recent additions to the SDSC repertoire of resources—Voyager, which provides AI-focused hardware for exploring AI in science and engineering, and the National Research Platform, which is a distributed infrastructure with compute hardware on the West coast, Midwest and East coast, and a content delivery system with caches in the national network backbone of Internet2 in five additional locations.

Mahidhar Tatineni, manager of the DESC User Services group said, "Across all of our supercomputers there are about 200,000 cores and about 1,500 accelerators, which deliver nearly two billion core-hours and over 12 million accelerator hours per year, and provide 46 petabytes of usable storage. About 100,000 users have been using these resources in recent years."

DESC members provide systems expertise in innovative systems management and significant domain and computational science expertise in using these machines. "In addition, DESC staff collaborate with many researchers in various capacities and run their own research programs," said Andreas Goetz, who leads the Computational Chemistry Group.

Throughout each year, DESC experts host numerous training sessions and workshops, including hackathons—design races in which computer programmers and others involved in software development work in small teams to create and/or optimize functioning software within a limited timeframe.

DESC also hosts SDSC's long-running Summer Institute featuring its members and others as speakers during the week-long supercomputing and data science training session. Other experiential learning projects include the Supercomputing Student Cluster Competition (SCC) for UC San Diego students, which is an international competition to build a cluster and run optimized applications to achieve the highest performance possible on an architecture. "SDSC also hosts a 14-week HPC user training series ... with focus on undergraduates, graduates, and the general research communities," said Mary Thomas, who leads the HPC training programs and directs the SCC.

"We have NSF grants related to CI training focusing on AI and neuroscience. Another program, running over a decade and managed by Ange Mason of DESC, is the Research Experience for High School Students (REHS) program, which allows high school students to do an eight-week long summer internship with mentors from various divisions of SDSC," said Bob Sinkovits, SDSC's HPC Education, Outreach and Training lead. "This is a tremendous program that allows high school students to gain important research experience and an environment to collaborate with others. These high school students have gone to top universities in the U.S., including UC San Diego."

DESC's HPC@MSI program is specifically targeted for Minority Serving Institutions (MSI) and focused on introducing them to national cyberinfrastructure resources.

"In addition to topics of supercomputing and data science, we host training sessions, on topics of science gateways for faculties from Hispanic Serving Institutions (HSIs) and collaborate with MSI faculties in various capacities," said Nicole Wolter, who manages the HPC@MSI and HPC@UC programs among others. "Many of our NSF-funded projects include strong efforts in Broadening Participation in Computing."

Broadening Participation in Computing
www.nsf.gov/cise/bpc/

EXTERNAL RELATIONS HAS SDSC COVERED

The External Relations (ER) Division at SDSC works to represent SDSC through various modes of communication—from news and feature stories, to social media and multimedia, to programs and events and web services. With the center's volume and variety of activities, the team of eight keeps busy with its comprehensive practices.

"The ER Division functions as the public relations and marketing arm of SDSC," said ER Division Director Cynthia Dillon. "Collectively, we offer a full set of communications tools which our highly skilled team members use to generate public attention for and engagement with the center."

The lively division is divided into four teams. The news team is made up of two science writers/editors who gather information about the research, partnerships, collaborations and other activities in which the SDSC community participates. The writers distill the academic and technical language into popular vernacular as much as possible to create a narrative adapted for a broad audience. The writers also develop and publish feature stories about SDSC PIs, researchers, partners and other stakeholders engaging in newsworthy activities. This content is shared across SDSC internal channels—website, social media and Innovators newsletter—as well as external outlets such as UC San Diego Today, national news wires, and media outreach.

"In addition to sharing information with the UC San Diego community and the overall high-performance computing research community, we often reach out to public affairs officers at funding agencies to ensure that they're in 'the know' about the great discoveries made possible by the amazing researchers at our center," said Kimberly Mann Bruch, science writer. "We also mentor secondary and university students each year as we feel helping to educate the next generation of science communicators is an important aspect of our job."

Within the news branch of the division is the ER editorial team, made up of the writers and the multimedia and website leads. Editorial strategy and content placement fall under the purview of this subgroup. The members work together on all aspects of news content and releases, the SDSC's Innovators newsletter, the annual report, the website, branding, marketing and other communications resources.

The multimedia team is made up of a full-time graphic designer and user experience lead, as well as a full-time social media and multimedia lead. This team is responsible for the graphics, photography, videography, social media and content templates used to publish SDSC's news and information content. This team also monitors and measures SDSC news and information engagement across communications channels.

"The ER group amplifies SDSC's activities and achievements. Our creative arm of the group, working closely with ER's science writers, attracts attention to help communicate complex ideas using clear and effective visuals in support of the center's brand. That begins with knowing the organization, the contributors and the target audiences," said Creative Lead Ben Tolo. "To create and deliver positive user experiences, we look to where audiences receive information whether online or in print, then engage them with interesting and useful content presented in a way that's easy to understand and appreciate."

The programs and events team includes a full-time programs manager and two full-time events specialists. Together they develop, plan and implement numerous conferences, training programs and workshops throughout each academic year. They also host SDSC's two summer institutes. Additionally, they lend support to SDSC's Industry Partnerships program and its Education & Training program.

"Our team works closely with programs across the center and external stakeholders to extend training, education and collaborative opportunities throughout the research community. Our programs and events are developed and hosted to support college to career professionals," said Program and Events Manager Susan Rathbun.

Web architecture and web services rounds out the division with a lead systems administrator who services multiple clients with their various website needs. The web services lead also supports the multimedia team and is part of the editorial team.

"We provide design, programming, technical and administrative capabilities for the SDSC website and CMS, along with similar services to dozens of SDSC and UC San Diego researchers, through our web hosting platform," said Web Services Lead Michael Dwyer. "We work closely with site owners to customize services to their needs and simplify processes and content management in support of SDSC's mission."

INDUSTRY RELATIONS CREATES IMPACT WITH TRIED AND TRUE TOOLKIT

The San Diego Supercomputer Center's Industry Relations (IR) program functions as a service provider to the center. While related to the External Relations (ER) Division, IR operates independently under the leadership of a director who receives support from two ER staff members who develop, plan and implement industry-related programs and events.

"Tell us how you want to partner with industry, and we will help structure a program to do that," said newly retired, long-time IR Director Ron Hawkins. "I view us as having developed a 'toolkit,' which includes sponsored research, focused centers of excellence, service agreements, gifts, etc., for working with industry. We deploy those tools as appropriate to develop the partnerships and collaborations that groups are pursuing."

According to Hawkins, IR contributes to SDSC's reputation as a national leader in high-performance and data-intensive computing, and cyberinfrastructure by developing industrial collaborations that practically demonstrate the value of HPC/CI to industrial research and development.

"We work with External Relations on communications that showcase SDSC's industry partnerships and their impact on advancing commercial R&D, technology transfer, economic impact and other aspects," said Hawkins.

Examples of IR's impact include its work with the High-Performance Wireless Research Network (HPWREN). When NSF funding for HPWREN expired, IR developed an innovative, user-community funding model that permitted this important research infrastructure to continue operating for more than 10 years following the end of NSF funding. In fact, HPWREN operations continue to this day.

"This effort exemplified the transition of NSF-funded basic research to a sustainable, operational system using a public-private partnership model," noted Hawkins.

Another example of IR's impact is its efforts toward contributing to the Human Genomics Revolution. Starting with its "data-intensive" Gordon supercomputer in 2011 and continuing to the present day, SDSC has supported numerous non-profit and for-profit entities in storing and analyzing the vast amounts of human genome data gathered by virtue of the revolution in Next Generation Sequencing (NGS) technology.

"The support provided by SDSC, including computing, storage and bioinformatics programming expertise, has contributed to multiple efforts leading to greater understanding of human genomics and practical applications for the diagnosis and treatment of disease," said Hawkins.

According to Hawkins, the present-day IR group was established in late 2007 when the center had few industry partnerships, and when researchers were frustrated by the "friction" of then existing mechanisms for working with industry.

"The newly established IR Department set about creating a toolkit of programs to facilitate working with industry and communications programs to attract potential partners. As a result of some successes over the years, I believe we have demonstrated the value of industrial partnerships and developed enthusiasm across the entire organization for pursuing industrial collaborations," said Hawkins.

RESEARCH DATA SERVICES DIVISION SERVES THE RESEARCH COMMUNITY WITH HEART

If high-performance and data-intensive computing and cyberinfrastructure make up the soul of the San Diego Supercomputer Center, then the Research and Data Services Division might just be at its heart. Known as RDS to the SDSC community, this division provides services that enable researchers to attain their research and computing goals.

According to Brian Balderston, director of infrastructure for RDS, the division's services include foundational needs for researchers—power, network and systems—as well as systems integration support.

"Our services also serve more variable needs of researchers, such as cloud computing, storage for active workloads or archival use cases, as well as backup storage for disaster recovery. We have the expertise to guide PIs to their research goals and have cultivated a vast network to grow partnerships more broadly," said Balderston.

RDS collaborates with researchers to identify, build and serve their research computing and data needs, which include compute, storage maintenance and research support. The division also offers on-premise and public cloud computing resources and solutions, often tailor-made to academic research needs. It offers on-premise services in the 19,000-square-foot, 3.5 MW (with potential capability of 13 MW), high-speed, network-connected data center. RDS also supports education with year-round internship opportunities for undergraduate students interested in software development, project management and other research computing experiences.

According to Christine Kirkpatrick, director of the division, much of what happens in RDS is only seen when things go wrong.

"We quietly work behind the scenes to anticipate what infrastructure will be needed for tomorrow's science and to deliver research computing services with a high degree of up time and good customer service to researchers," said Kirkpatrick. "All activities, especially the Data Center operations and the Help Desk, carried on at full speed even with the uncertainty of our times. Our infrastructure teams, led by Brian Balderston, have been fortunate to grow during this period, due in large part to the spirit of fun and collegiality alive in RDS, as well as the flexibility afforded by leadership to retain and attract top talent."

Kirkpatrick explained that the enterprise networking team is upgrading RDS' backbone to 400Gb capacity. The storage service, USS/Qumulo, continues to be a runaway hit with contracts for 20 petabytes and growing since its inception. Additionally, the platforms and cloud integration teams continue to deliver excellent service to individual researchers, high-profile research partners in our region and internationally, UC San Diego departments and other UC partners.

Over the past few years, RDS has been steadily working on the intersection of artificial intelligence (AI)/machine learning (ML) and FAIR Principles (findable, accessible, interoperable and reusable research objects including data), as well as reproducibility.

"My own research is in data-centric AI, working at the intersection of ML and FAIR, with a focus on making AI more efficient to save time and power consumption—for costs and carbon footprint concerns. Our Senior Cloud Integration Engineer Kevin Coakley conducts research related to AI reproducibility," said Kirkpatrick. "It can be easy to focus on employing techniques like ML, but many people don't realize that ML processes may need to be run multiple times and that the results can vary between laboratories (the term that wraps up everything about a specific processing environment including the hardware and software versions). These differences in results can sometimes change the scientific inference meaning that is taken away. The Open Science Grid, led by our SDSC Director Frank Würthwein has been a tremendous resource for re-running ML processes on several different types of clusters."

Another NSF-funded initiative co-led by RDS is the West Big Data Innovation Hub, which aims to build and strengthen partnerships across academia, industry, nonprofits and government—connecting research, education and practice to harness the data revolution. Most recently, RDS staff, led by Kim Bruch, worked with the Pala Native American Youth Council on a national DataJam project and was awarded "Best New Team" at the final competition.

Other recent RDS accomplishments include the recent replacement of 20,000 pounds of toxic lead-acid batteries with a safer, environmentally friendly and cost-effective alternative. The project in partnership with Urban Electric Power will more than double available battery backup electricity.

"SDSC is the world's first enterprise application of this innovative rechargeable battery technology, and our partnership with Urban Electric Power has made our computing footprint greener," said Kirkpatrick in a previous news article (April 20, 2022; SDSC/UC San Diego).

According to Balderston, RDS serves several of the core principles of UC San Diego—primarily research and education.

"Our work contributes to discoveries in the cosmos, under the oceans and novel healthcare advances. It supports efforts to combat natural disasters and to apply research for social good. And it serves the planet in the fight against climate change," said Balderston. "We are also working to ensure that research efforts are achieved in an equitable and ultimately FAIR fashion. We provide stable services and functional structures that researchers can count on."

RDS includes experts in the following areas: platform services, cloud and storage, enterprise network services, help desk, operations, research data initiatives, project management and student interns.

Members of SDSC's Research and Data Services Division.

NSF's Steven Ellis offers support options for science gateway developers at the Gateways 2022 conference in San Diego.

SUSTAINABLE SCIENTIFIC SOFTWARE DIVISION MAKES COMPLEX SOFTWARE AND COMPUTING RESOURCES EASY

Creating and managing the computational resources necessary for modern science is a challenge that is faced by all branches of science. To enable scientists to concentrate on their research needs, SDSC's Sustainable Scientific Software Division (S3D), led by Director Michael Zentner, focuses on providing several areas of services:

1. **cyberinfrastructure**—using science gateway platforms to make scientific software and hardware widely available to everyone;
2. **software and project sustainability planning**—exploration of revenue-generating opportunities as an alternate means of supporting the ongoing development of scientific software;
3. **professionalized software development and operations services**; and
4. **next generation tools for biology**—development and operation of the Cyberinfrastructure for Phylogenetic Research (CIPRES) science gateway.

"Example programs within our division include SDx, which encompasses professional software development and operations, and Rev-Up, which helps our clients work toward revenue generation strategies for sustainability," said Amit Chourasia, S3D's associate director and Director of S3D's Hubzero® project. "We also have a strong bioinformatics team, including our CIPRES science gateway framework team and last but not least my work that revolves around new developments for the Hubzero platform."

Hubzero is an open source software platform for building powerful science gateways that host analytical tools, publish data, share resources and enable users to collaborate and build communities. Initially created by researchers in the National Science Foundation-sponsored Network for Computational Nanotechnology to support nanoHUB.org, Hubzero now supports numerous additional projects including NEMAR.org, QUBEShub.org and Geodynamics.org. S3D staff go beyond typical platform hosting and software development—they also provide sound advice for clients related to new audience acquisition and project sustenance.

"With over two million visitors coming to the over 20 Hubzero-operated science gateways each year, the academic community can access open and citable research products shared on these science gateways," said associate director and lead of S3D's SDx team Rich Wellner. "SDx strives to not only provide our users with solid software development and cyberinfrastructure operations, but also allows them to utilize our knowledge and skillsets in developing their product and ensuring it is successful for their audiences. We are also provide software development and operations services for the Hubzero platform itself. This allows us to use our engagements with users in a variety of scientific disciplines to enhance and expand Hubzero."

The Hubzero platform operates for communities from a wide variety of disciplines—from community-driven geospatial data modeling to online plant science collaboration for K-12 students to open access to ice-sheet datasets, tools and resources that improve estimates of future sea level rise. All of these projects share the same foundational platform and have used the platform's features to enable their communities to publish research products, run software tools, collaborate with colleagues and stakeholders, visualize data models and build interactive educational resources.

"We have developed Hubzero as a platform where people can analyze datasets and use interactive simulation tools via modern web applications such as RStudio, Jupyter and Notebooks—also through traditional compiled applications like C/Fortran codes and MatLab," explained Chourasia.

According to Wellner, Hubzero also helps users publish research products including datasets, software tools and white papers through a step-by-step, guided system—as well as provide a place for collaborators to discuss their work, track progress and share digital assets.

"Through SDx, we provide robust 24/7 'no-hassle' operations for Hubzero-based science gateways so researchers can focus on their science rather than IT issues," said Wellner.

In addition to Hubzero, SDSC's S3D is also home to the Science Gateways Community Institute (SGCI) where community members benefit from shared, open source research and education resources from around the world. SGCI members also receive software development support, usability support and more. Since 2016, 170 science gateways have received comprehensive design, development and sustainability consulting to enhance their projects through the institute. According to SGCI, clients, on average, have estimated that SGCI accelerated their efforts by more than seven times.

S3's SGCI also runs educational programs related to science gateway development and sustainability. Over 660 students and faculty from underrepresented groups have received support since the institute's inception.

According to Zentner, SGCI has written more than 220 letters of commitment since inception, with 74 awarded. From those awarded, science gateway projects have designated $3 million to SGCI.

"The team of people who all work to help each other and are motivated by their contributions to science and society is what makes the S3D rewarding for me," said Zentner. "We continuously support research and education for so many people worldwide with our science gateways. We have also created new ways of thinking about projects through Rev-Up and SGCI. Our work changes people's mental models of how their projects might progress in the future."

NSF AWARDS NEW SGX3 CENTER OF EXCELLENCE TO SDSC

In 2022, the NSF awarded a new Center of Excellence to Extend Access, Expand the Community, and Exemplify Good Practices for CI through Science Gateways (SGX3). SGX3 will enhance many of the services of the SGCI by increasing its focus on bringing more domain scientists into the cyberinfrastructure community and increasing the awareness of how science gateways can serve these various science communities. Importantly, SGX3 will include forward looking activities to define roadmaps of new technological advances that will be needed over the next 5-10 years for science gateways to serve emerging scientific needs. This will be particularly important as AI and the different models of computing it introduces become part of nearly all research domains, and to increase the breadth of usage for the heterogeneous NSF-funded collection of compute resources to those who would otherwise need to develop advanced skills to use such resources.

CORE VALUES SHAPE SUCCESS OF SHERLOCK TEAM TO DELIVER CYBERSECURITY RESULTS

Similar to a famous fictional "Holmes" who doggedly works to crack a case, the Sherlock Cloud Solutions and Services Division (Sherlock) at SDSC works unwaveringly to solve the mysteries that cyberinfrastructure and cloud computing can present to the people and places the division serves.

With its unique array of experts who apply their own set of keen skills to the services and solutions the division has to offer, "Sherlock" contributes to the professional, research and scientific aims of stakeholders such as the SDSC community, the University of California, external academic and research institutions, and industry partners.

"At a high level, our core expertise is at the intersection of cloud computing, cybersecurity and regulatory compliance," said Sandeep Chandra, executive director of the division. "To support these key areas, we have organized ourselves across groups within the division."

These groups include:

- Cloud Architecture and DevOps
- Cloud Infrastructure
- Data Architecture and Platforms
- Cybersecurity and Compliance
- Outreach and User Support

Notably, Sherlock currently partners with the UC San Diego Information Technology Services (ITS) to develop and deploy a Cybersecurity Maturity Model Certification (CMMC) hosting solution that will span across Microsoft Azure Gov Cloud and Office365 GCC High.

"This solution is first of its kind within the UC system and will serve research organizations on campus, other UC campuses and institutions across the country," noted Chandra.

Sherlock also works extensively with a variety of federal agencies including, but not limited to, the Centers for Medicaid and Medicare Services (CMS), National Institutes of Health (NIH), National Science Foundation (NSF) and the Department of Defense (DoD). Through its work with these agencies, Sherlock has deployed numerous secure, compliant, end-to-end cyberinfrastructure solutions to support many critical applications such as cancer research, decoding the human immune system, and detecting and preventing medical fraud.

Leslie Morsek, who leads Sherlock's outreach efforts and contributes to program development, explained that the division operates the technologies, applications and underlying platforms with agility while mindful of inventive and modern solutions to enable growth with ever-changing technological, regulatory and unique partner requirements. Sherlock couples this operation with ensuring data security and privacy.

According to Winston Armstrong, who leads the Security and Compliance Group, Sherlock has contributed to SDSC's reputation as a national leader in high-performance and data-intensive computing and cyberinfrastructure in numerous ways.

"Sherlock is rooted in cloud, cybersecurity and regulatory compliance knowledge and expertise; it built and deployed its Sherlock Cloud, which offers HIPAA-, FISMA- and NIST 800-171 Controlled Unclassified Information (CUI)-compliant compute, data management and application hosting services," said Armstrong. "The Sherlock Division further strategized and deployed a multi-cloud solution that incorporates Amazon Web Services (AWS), Microsoft Azure (Azure) and Google Cloud Platform (GCP) to provide partners with a path to public cloud adoption and migration for regulatory compliance workloads."

Eric Odell, one of Sherlock's senior cloud architects, pointed out that Sherlock partnered with UC San Diego's ITS to build a secure environment in Azure Government Cloud to support new regulatory requirements coming out of the U.S. Department of Defense.

"This new capability will position UC San Diego to be a leader within the UC system and academia nationwide," he said.

Morsek noted that as a result of its innovation and exceptional solutions, Sherlock is a UC Center for Excellence in Regulatory Data Management Services.

"Moreover, it is one of the few academic organizations that operates this capability as a service for other academic, government, research and industry partners," she said.

While Sherlock is committed to excellence and innovation in providing partners with highly secure, compliant, versatile and successful solutions to meet their research needs, the division recognizes the importance of its intangibles—namely, its team members and the customer service it provides.

"The Sherlock team is comprised of individuals with varied expertise and talent and each possess an unwavering commitment to make an impactful contribution to Sherlock, its partners and those who indirectly benefit from the solutions deployed. This teamwork and our customer-serving mindset is extraordinary and has led to the success of Sherlock and its partners," said Chandra, adding, "At Sherlock, we value purpose, mastery and autonomy. We are proud that everyone in our team gets to experience these core values."

RECENT EXAMPLES OF SHERLOCK'S MOST IMPACTFUL WORK

UCOP RISK DATA MANAGEMENT SYSTEM: The Risk Services Data Management System (RDMS) is a reporting application that equips staff at the UC campuses and health centers with the data that helps them make data-driven decisions to reduce the overall cost and impact of risk. RDMS enables UC systemwide staff to identify and develop strategies to minimize the impact of risk, assess the effectiveness of safety programs, ensure the highest quality of care and patient safety across the UC health system, and monitor claims and claim costs. Sherlock has had a long-standing partnership with UCOP Risk Services and has helped implement innovative solutions for their stakeholders over the years.

TEMPREDICT: Sherlock partnered with researchers from the University of California, UC San Francisco, UC San Diego, Massachusetts Institute of Technology (MIT) Lincoln Laboratory, the U.S. Army and the U.S. Navy to deploy a HIPAA-compliant platform in AWS to collect physiological data from frontline healthcare workers and the general population. The aim of the study was to complete antibody testing for 10,000 participants and to provide additional support for algorithm development and testing in real-world settings. Sherlock partnered with these organizations and external vendors to build cyberinfrastructure for secure storage for streaming data, data processing leveraging time series databases, and analytical tools for the TemPredict researcher community.

CALIFORNIA TEACHERS STUDY (CTS): Sherlock partnered with the City of Hope, a National Cancer Institute (NCI) designated Comprehensive Cancer Center, to develop and deploy a research cyberinfrastructure that included a secure, cloud-based data management and analytics platform. This platform allowed every member of the CTS team to securely access and use all CTS data and information in real-time in a consolidated, integrated, and secure manner. It also securely integrated in real time with other public Cloud platforms. Working cooperatively with City of Hope leadership, Sherlock developed innovative data management and analytics solutions that have transformed and modernized the way in which City of Hope secured and protected its data and enabled research.

MEDICAID PROGRAM INTEGRITY: Sherlock's collaboration with the Centers for Medicare and Medicaid Services (CMS), part of the federal Department of Health and Human Services, gave rise to a Federal Information Security Modernization Act (FISMA) certified, high-performance data warehouse and analytics platform to identify instances of Medicaid fraud, waste and abuse. The system enabled review of actions by individuals or entities furnishing medical products or services and claiming reimbursement through the Medicaid program. The system stored provider, claims and referential data from all state Medicaid agencies and provided rich data mining and analysis software tools to enable the federal Medicaid Integrity Program.

Centers of Excellence

The Centers of Excellence at SDSC are part of a larger strategic focus to help researchers across all domains—including those who are relatively new to computational science—better manage the ever-increasing volume of digitally based information. These centers formalize key elements of SDSC's wide range of expertise, from big data management to the analysis and advancement of the internet. Below is information about each of these impactful centers.

CENTER FOR APPLIED INTERNET DATA ANALYSIS (CAIDA)

Formed in 1997, CAIDA is a collaborative undertaking among organizations in the commercial, government and research sectors. It is aimed at promoting greater cooperation in the engineering and maintenance of a robust, scalable global internet infrastructure.

Kimberly "kc" Claffy, CAIDA's director, principal investigator and co-founder, has been a leader and pioneer in internet science for nearly three decades. Inducted into the Internet Hall of Fame in 2018, Claffy, along with Hervey Allen of the University of Oregon's Network Startup Resource Center (NSRC) and David Clark of MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL), was awarded more than $11 million by the National Science Foundation (NSF) for two projects aimed at improving internet infrastructure security. Claffy leads the projects with a team of U.S. and international collaborators. The first project, Global Measurement Infrastructure to Improve Internet Security (GMI3S), is aimed at supporting the design and prototyping of a distributed but integrated infrastructure to measure internet topology and traffic dynamics, with the intention of improving internet infrastructure security. The goal of the second project, Integrated Library for Advancing Network Data Science (ILANDS), is to understand the internet's changing character through realistic datasets and longitudinal measurements, as well as new experiments with accessible data for researchers.

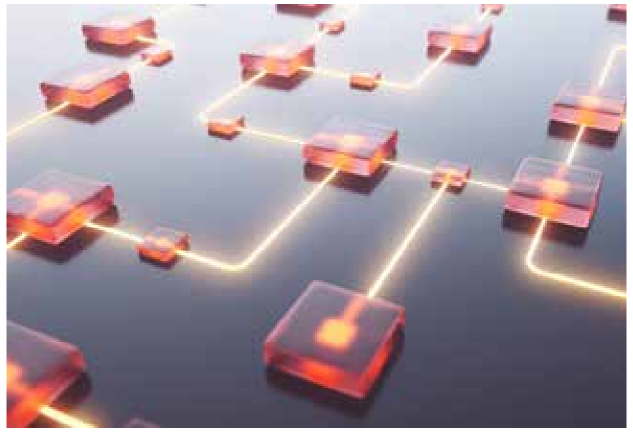
CENTER FOR LARGE-SCALE DATA SYSTEMS RESEARCH (CLDS)

CLDS was established in 2012 by SDSC's James Short and Chaitan Baru as an industry-university partnership to study and address technical as well as technology management-related challenges facing information-intensive organizations in the era of big data—essentially CLDS uses data to solve real-world problems. It achieves this by experimenting with data and technology to understand and solve complex business, economic and social challenges.

CLDS specializes in developing applicable concepts, frameworks, analytical approaches, case analyses and systems solutions to big data management, with a related goal of developing a set of benchmarks for providing objective measures of the effectiveness of hardware and software systems dealing with data-intensive applications. Based at SDSC to leverage the center's resources and large-scale compute and storage resources, CLDS operates from the premise that the data economy will not be built on data systems alone. Rather, data is changing the social and economic landscape, concentrating market power in a few industry sectors and diminishing the relative value in others. For individuals, data is raising concerns about trust, privacy and security, as well as the equitable distribution of gains from data. A market economy cannot function without trust. Trust deficits can unravel data markets and undermine civic and social cohesion. All of these factors must be considered in understanding and contributing to the evolving data economy. CLDS is organized into research program areas, connected laboratories, a unique industry-driven business exchange (BX) program, educational workshops and workforce education and training.

SHERLOCK

Sherlock debuted in November 2008, and launched its flagship service Sherlock Cloud, an Infrastructure as a Service (IaaS) capability that complied with the Federal Information System Management Act (FISMA), rendering it the largest FIMSA-certified cloud within the UC system. The cloud infrastructure was further developed in accordance with hundreds of National Institute of Standards and Technology (NIST) controls governing system access, information control and management processes; it also addressed federal Cloud First requirements. Sherlock Cloud Solutions and Services was established in 2013 to provide managed services to meet the secure computing and data management needs of our academic, government and industry customers. Its approach involves building the solution, managing the software and hardware platform, and the necessary management processes that govern it. Sherlock's comprehensive solution portfolio is a proven resource, and the division stands on the belief that its intangibles, namely its superb team, experience and knowledge, provide its customers with an edge. The team works with its customers to jointly solve problems and create an environment that not only meets, but surpasses, their needs. Directed by SDSC's Sandeep Chandra, Sherlock's capabilities range from compliance, cloud computing and cybersecurity to DevOps, unified data analytics and data management. Recently, Sherlock partnered with the University of California Office of the President's (UCOP) risk and technology delivery services groups and Kwartile to successfully re-architect and migrate the UCOP Risk Services Data Management System (RDMS 1.0) from an on-premise, Hadoop-based platform to a server-less, data lake platform in the Amazon Web Services (AWS) Cloud (RDMS 2.0).



WORKFLOWS FOR DATA SCIENCE CENTER (WORDS)

Formed in 2014 and housed at SDS, the Workflows for Data Science (WorDS) Center of Excellence is a hub for the development, promotion and delivery of workflow services for a wide range of applications. Its mission is to support data analysis projects, data scientists and software engineers in their computational practices involving process management. According to Director Ilkay Altintas, WorDS aims to assist researchers in creating workflows to better manage the tremendous amount of data being generated across a wide range of scientific disciplines—from natural sciences to marketing research—while letting them focus on their specific areas of research instead of having to solve workflow issues or the computational challenges that arise as data analysis progresses from task to task. WorDS expertise and services include consulting with world-class researchers and an A-Team of developers well-versed in data science and scientific computing technologies; workflow management technologies that resulted in the collaborative development of the popular Kepler Scientific Workflow System; development of data science workflow applications through a combination of tools, technologies and best practices; hands-on consulting on workflow technologies for big data and cloud systems, i.e., MapReduce, Hadoop, Yarn, Cascading; and technology briefings and classes on end-to-end support for data science. WorDS' researchers collaborate across a range of scientific domains that include: bioinformatics, environmental observatories, oceanography, computational chemistry, fusion and geoinformatics. WorDS is funded by a combination of sponsored agreements and recharge services.

Physics, Computation Experts Help Earn $15M to Advance AI, Data Analysis

As scientific data sets become progressively larger, algorithms to process the data become more complex. Artificial Intelligence (AI) has emerged as a solution to efficiently analyze these massive data sets, and new computer processor types—such as graphics processing units (GPUs) and field-programmable gate arrays (FPGAs)—help speed up the work of AI algorithms. This combination of AI and new processor types is leading to a revolution in the realm of data analysis.

In an effort to shift direction in the application of real-time AI at scale, the National Science Foundation (NSF) funded $15 million in support of advancing scientific knowledge and discovery with the Accelerated AI Algorithms for Data-Driven Discovery (A3D3) Institute. Its mission is to incorporate AI algorithms with new processors to support analyses of these unprecedented data sets.

“AI-assisted analysis of multidisciplinary data sets will be critical in helping researchers locate and explore trends that can lead to new discoveries,” said UC San Diego Chancellor Pradeep K. Khosla. “The new multi-disciplinary and geographically distributed A3D3 Institute, supported through NSF’s Harnessing the Data Revolution (HDR) program, will lead the way with a collaborative team of researchers from UC San Diego, Caltech, Duke University, MIT, Purdue University, UIUC, University of Minnesota, University of Washington and the University of Wisconsin-Madison.”

UC San Diego influence

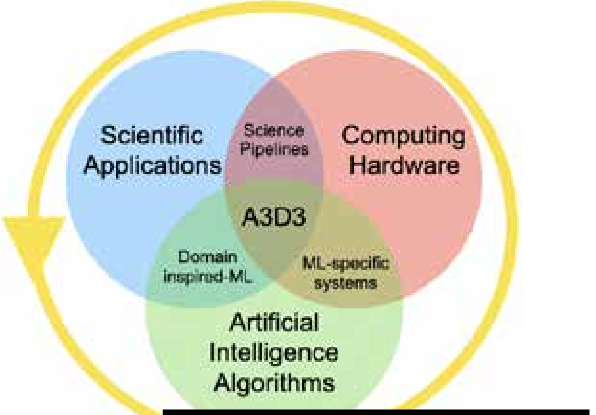
UC San Diego’s Javier Duarte, assistant professor in the Department of Physics who collaborates with researchers at SDSC, is an institute principal investigator (PI) and the university’s A3D3 board representative, as well as the equity and career representative on the executive board. In that capacity, he also co-supervises the post-baccalaureate program with Mia Liu at Purdue University. Frank Würthwein, SDSC director, participates, as does Amit Majumdar, who leads SDSC’s Data Enabled Scientific Computing Division. Additionally, UC San Diego postdoctoral researcher Daniel Diaz and graduate student researchers Raghav Kansal, Farouk Mokhtar and Anthony Aportela are developing accelerated AI algorithms.

Duarte said that his work is split between developing ultrafast machine learning algorithms deployed in specialized hardware, such as FPGAs, that can be used to process data from sensors in real time, and developing heterogeneous computing pipelines to enable faster processing of big scientific data.

Targeting fields of science for fast AI and outreach

To take full advantage of fast AI, the A3D3 Institute targets fundamental problems in three fields of science: high energy physics, multi-messenger astrophysics and systems neuroscience.

“A3D3 works closely within these domains to develop customized AI solutions to process large datasets in real-time, significantly enhancing their discovery potential,” said Duarte. “The ultimate goal of A3D3 is to construct the institutional knowledge essential for real-time applications of AI in any scientific field.”



Credit: Assistant Professor Philip Harris, MIT

Duarte also noted that A3D3 empowers scientists with new tools to deal with the coming data deluge through dedicated outreach efforts.

“The post-baccalaureate program, for example, is specifically aimed at helping underrepresented minority students, identifying as Black, Latinx, Indigenous, women and/or LGBT+, from institutions without extensive research opportunities, gain valuable research experience in order to ‘bridge the gap’ between undergraduate and graduate programs,” he said.

The director and PI of the A3D3 Institute is the University of Washington’s Shih-Chieh Hsu, a colleague and former student of Würthwein’s, who gave an example of the potential impact of the work at the institute.

“At the Large Hadron Collider (LHC), the challenge of processing data is daunting. With future aggregate data rates exceeding one petabit per second, the data rates at the LHC exceed all other devices in the world,” Hsu explained. “The aim of A3D3 is to build a series of tools that will enable the processing of all of this information in real-time using AI. Through the use of AI, A3D3 aims to perform advanced analyses, such as anomaly detection, and particle reconstruction on all collisions happening 40 million times per second!”

Real-time analyses in astrophysics and neuroscience

For UC San Diego-based projects out of SDSC, such as Voyager, an experimental AI research resource, and the Prototype National Research Platform (PNRP), a first-of-its kind testbed for a cyberinfrastructure ecosystem, Majumdar said there is a lot of synergy between them and the institute, given A3D3’s incorporation of AI algorithms and new processors.

“Voyager is based on dedicated AI hardware from Habana, while PNRP includes both FPGAs and more conventional GPUs,” noted Würthwein.

Duarte noted that within the field of multi-messenger astrophysics, A3D3 will be working to integrate AI to promptly and computationally process the data from telescopes, neutrino detectors, and gravitational-wave detectors efficiently in order to quickly identify astronomical events corresponding to the most violent phenomena in the cosmos.

“The ability to identify and further distribute these events as astronomical alerts enables the entire transient astronomy community to cross-correlate observations and understand astrophysical phenomena across multiple different forces,” Duarte said.

Amy Orsborn, an assistant professor in the Department of Electrical and Computer Engineering and the Department of Bioengineering at the University of Washington, explained that in systems neuroscience, A3D3 is working to discover the computations that brain-wide neural networks perform to process sensory and motor information during behavior. To do so, A3D3 will develop and implement high-throughput and low-latency AI algorithms to process, organize and analyze massive neural datasets in real time.

“These real-time analyses will enable new approaches to probing brain function such as causal, closed-loop manipulations. Applying powerful AI methods to systems neuroscience will significantly advance our ability to analyze and interpret neural activity and its relationship to behavior,” said Orsborn.

According to the NSF, institutes such as A3D3 will enable will enable breakthroughs through collaborative, co-designed programs to formulate innovative data-intensive approaches for addressing critical national challenges. First outcomes are expected in 2023.

This project is supported by the NSF (grant no. 2117997).

Physicists Apply FAIRness to Data Studies

For scientists in observational disciplines, data is the lifeblood of research. Collecting, organizing and sharing data both within and across fields drives pivotal discoveries that benefit society and help make it more secure.

Making data open and available, however, is only part of the answer to the question of how different scientists—often with very different training—can draw useful conclusions from the same dataset. In order to promote and guide the cultivation and exchange of data, researchers have developed a set of principles that could make the data more findable, accessible, interoperable and reusable (FAIR) for both people and machines. And SDSC is leading national efforts to align principles with practice.

Although the FAIR principles were first published in 2016, researchers are still figuring out how they apply to particular datasets. In a new study, researchers from UC San Diego, the U.S. Department of Energy's Argonne National Laboratory, the Massachusetts Institute of Technology and the Universities of Minnesota and Illinois Urbana-Champaign have laid out a set of new practices to guide the curation of high-energy physics datasets that make them more FAIR.

The research demonstrates how to FAIRify an open simulation dataset, consisting of Higgs boson decays and quark and gluon background, produced by the CMS Collaboration at the CERN Large Hadron Collider (LHC).

"The dataset is extremely complex even for expert particle physicists, so a major question related to FAIRness we sought to address was how to convey the necessary information even to nonexperts," said Javier Duarte, CMS collaborator and assistant professor of physics at UC San Diego. "To really enable the reusability of the massive datasets that will be produced by the LHC and other experiments, we have to ensure any scientist can understand the data."

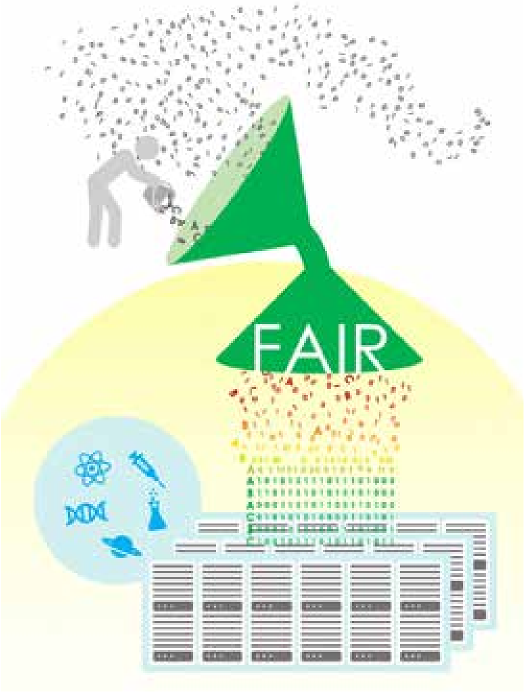
The production of FAIR data and other digital objects has become a powerful notion throughout the research world, aimed at increasing successful data integration and allowing for seamless service provision across multiple resources and organizations. SDSC Research Data Services Division Director Christine Kirkpatrick serves as a leader in FAIR data efforts via the U.S. GO FAIR Office, led out of SDSC.

"FAIR focuses attention on the need to more closely align research data management practices towards machine actionable data, code, workflows, AI models and other digital objects," explained Kirkpatrick.

To assist researchers from other domains and highlight the interplay between AI research and scientific visualization, the recent study also provided software tools to visualize and explore this FAIR dataset.

"The FAIR principles were created to serve as goals for data producers and publishers to improve data management and stewardship practices," said Argonne computational scientist Eliu Huerta. "The community expects that adhering to these principles will enhance the capabilities of machines to automate the finding and use of data, thereby streamlining the reuse of data for humans."

In addition to building FAIR datasets, the research team also sought to understand the FAIRness of AI models. "To have a FAIR AI model, we believe you need to have a FAIR dataset to train it on," said Yifan Chen, a graduate student at the University of Illinois Urbana-Champaign and Argonne's Data Science and Learning division. "Applying the FAIR principles to AI models will automate and streamline the design and use of those models for scientific discovery."



FAIR AI models and data may be coupled with modern scientific data infrastructure and innovative computing to automate and accelerate discovery. Credit: Argonne Leadership Computing Facility Visualization and Data Analytics Group.

"Our goal is to shed new light into the interplay of AI models and experimental data and help create a rigorous framework for the development of AI tools to address the biggest challenges in science," Huerta added.

"For the first five years, the focus of FAIR was on data. The conversation and practices have now moved on to making all aspects of data and computationally intensive research FAIR including workflows, science gateways, software and AI models," said Christine Kirkpatrick, head of the GO FAIR U.S. Office and division director of Research Data Services at SDSC.

Ultimately, Huerta said, the goal of FAIRness is to create an agreed-upon set of best practices and methodologies, which will maximize the impact of AI and pave the way for the development of next-generation AI tools. "We're looking at the entire discovery cycle, from data production and curation, design and deployment of smart and modern computing environments and scientific data infrastructures, and the combination of these to create AI frameworks that power disruptive advances in our understanding of scientific phenomena," he said.

Supercomputers Provide Single Model of Inner and Outer Solar System

Supercomputer-enabled models generated at SDSC and the Pittsburgh Supercomputing Center (PSC) revealed new insights into the solar system's formation. While attempting to better understand the relationship between Earth and Mars, postdoctoral fellow Matt Clement (Carnegie Institution of Washington's Earth and Planets Laboratory) and his colleagues used allocations from the NSF's Extreme Science and Engineering Discovery Environment (XSEDE) on Comet at SDSC and Bridges at PSC to illustrate the formation of both the inner and outer solar system in a single model. Their work was published in the journal *Icarus*.

The international team included Clement, Nate Kaib (University of Oklahoma), Sean Raymond (University of Bordeaux) and John Chambers (Carnegie Earth and Planets Laboratory). Clement said that the team used numerical models to study the formation of the solar system's inner, terrestrial planets: Mercury, Venus, Earth and Mars.

"We know the giant planets—Jupiter, Saturn, Uranus and Neptune—must have formed far quicker than the terrestrial planets because observations of other forming solar systems in the galaxy indicate that the main ingredient for the gas giant planets, free gas, has been around for only a million years or so," Clement said. "Because isotopic dating of rocks from the Earth and moon give much later dates, the inner terrestrial planets must have formed in the presence of the fully grown giant planets, and we think this process occurred over a time span of around 100 million years, with the planets themselves slowly accreting and coalescing from a sea of small, asteroid-like objects."

This would have been a very violent time for the young Earth, as it would have been continuously smashing into other proto-planets and small objects en route to achieving its present size and orbit, according to Clement.

As for the outer solar system planets, Clement said that many peculiar aspects—such as swarms of asteroids that orbit along with Jupiter and Neptune and irregular moons such as Triton—are explained by the giant planets passing through a tumultuous epoch of instability at some point after their formation. During this instability, Clement explained that the planets' orbits evolved rapidly and substantially with their orbits becoming more elliptical and diverging from one another.

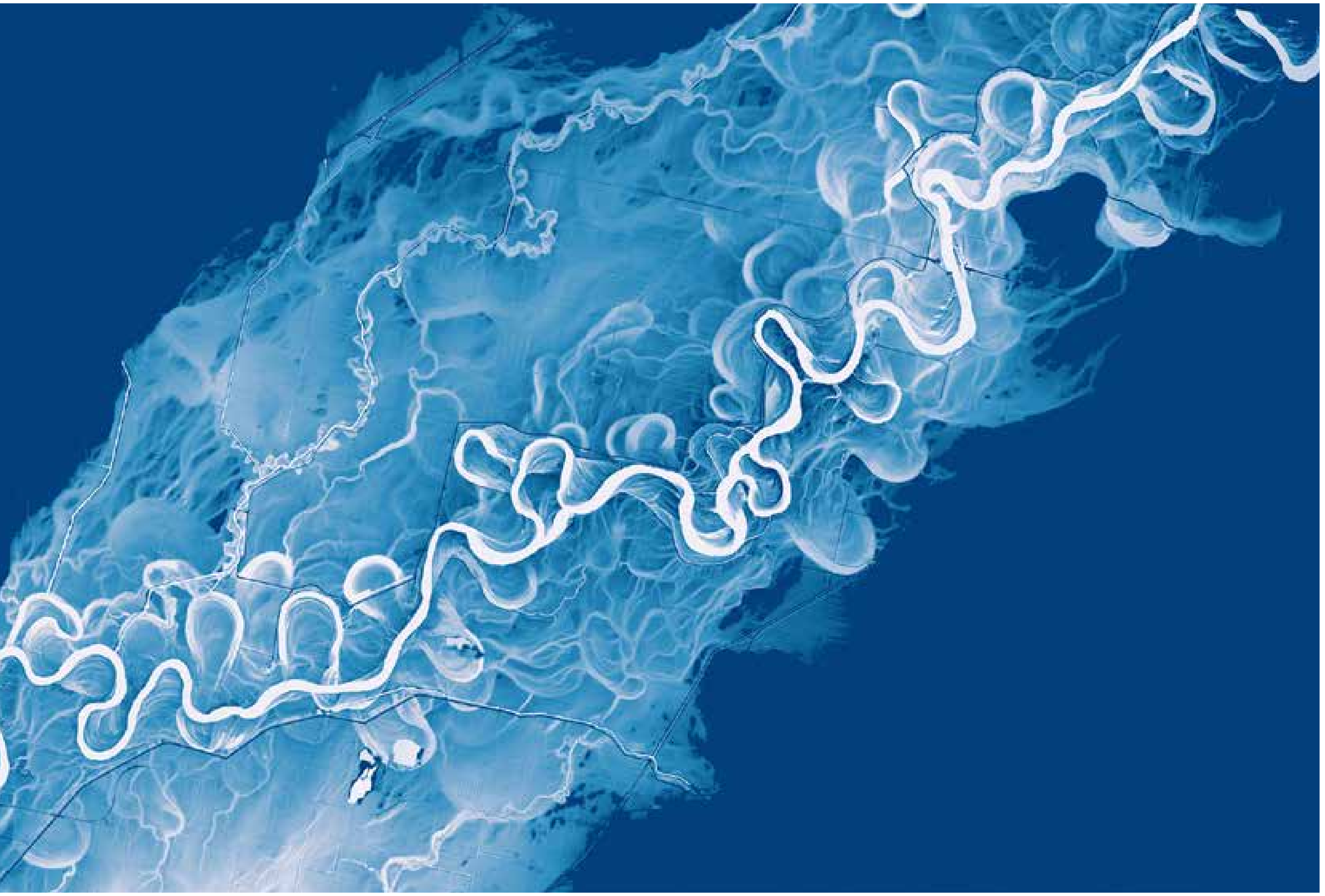
"It is suspected that planets similar to Uranus and Neptune once existed in between Saturn and Uranus, but they were ejected during this 'giant planet instability', which has become known as the Nice Model—as in Nice, France, where it was developed by scientists in the early 2000s," Clement said. "The Nice Model is arguably the consensus model for the formation of the outer solar system, however, it does not reconcile the orbits and masses of inner terrestrial worlds, and contemporary models of the inner solar system's formation fail to replicate the low mass of Mars, and its rapid geologically inferred formation time with respect to the Earth."

According to Clement, when the instability happens while the inner planets are still growing, the instability also explains why the inner solar system looks the way it does. "Thus, our work explains both the inner and outer solar system in a single model," said Clement.

More about Mars

While the primary objective of Clement and the team was to better understand the relationship between Earth and Mars, their study also revealed a potential resolution to another solar system mystery: Mercury's diminutive size and isolated orbit. That is, while Mercury is physically close to Venus, gravitationally speaking it is quite isolated as it makes nearly three revolutions around the sun for every one Venus cycle. The team's study found that, in certain simulations, interactions between the giant planets and the forming terrestrial material liberate a proto-planet from the Mars-region. That is, one of those objects that might have turned Mars into a bigger planet if the instability hadn't stunted its growth, has been implanted in the inner solar system on a Mercury-like orbit.

"If this genesis scenario is correct, it would have huge implications for our understanding of the terrestrial planets' compositions, as we would expect Mercury to be made of much the same material as Mars," Clement explained. "While Mercury's bulk composition is fairly unconstrained, forthcoming missions such as BepiColombo to the innermost planet should help improve our understanding of the planet's makeup and inform our models—we will continue to study this scenario on the new supercomputers Expanse at SDSC and Bridges-2 at PSC."



Researchers Produce First Map of Topographic Change at Statewide Scale

Scientists study the topography—the forms and features of the landscape—to measure and observe changes at the Earth's surface over time. While some changes are the result of natural processes like fluvial erosion and coastal erosion, the topography can also change due to anthropogenic forces, including those related to urban development, agriculture and resource extraction.

The OpenTopography team examined topographic change over the entire state of Indiana across the span of almost a decade. The scientists evaluated landscape changes driven by infrastructure development, vegetation growth, agricultural practices, river and coastal processes, and natural resource extraction.

Led by Arizona State University (ASU) researcher Chelsea Scott, in collaboration with researchers from SDSC and UNAVCO, the study included observations of physical changes to the Earth's surface such as movement and removal of rock material in quarries, erosion and deposition by major rivers, significant erosion along the Lake Michigan shoreline near Indiana Dunes National Park, new housing developments and highways, and land use changes related to agricultural activities.

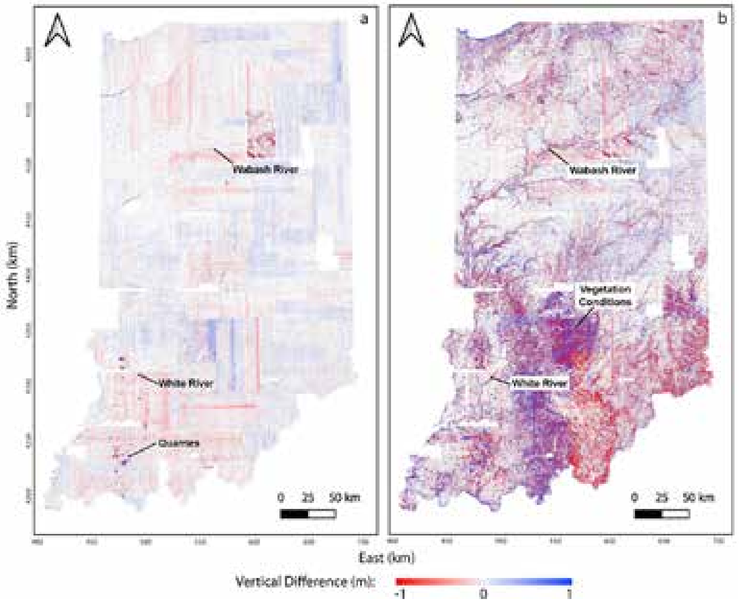
Using high-resolution lidar (light detection and ranging) topography data collected over the state of Indiana between 2011–2013 and 2016–2020, the researchers calculated both the change of the bare earth surface (minus the vegetation and built structures) and the full surface (the bare surface plus above-ground features like vegetation, buildings, and other structures).

OpenTopography already provides web-based, on-demand tools for its users to compute landscape change for its hosted spatially overlapping lidar datasets. These tools include vertical and 3D differencing algorithms as described in a research article in the journal *Geosphere*. The compute intensive nature of these algorithms and massive volumes associated with lidar data restrict processing to limited spatial extents at a time. Given the large volumes of data that had to be processed for this Indiana project, the team leveraged dedicated high-performance computing (HPC) resources at SDSC for faster processing and more reasonable timelines for analysis.

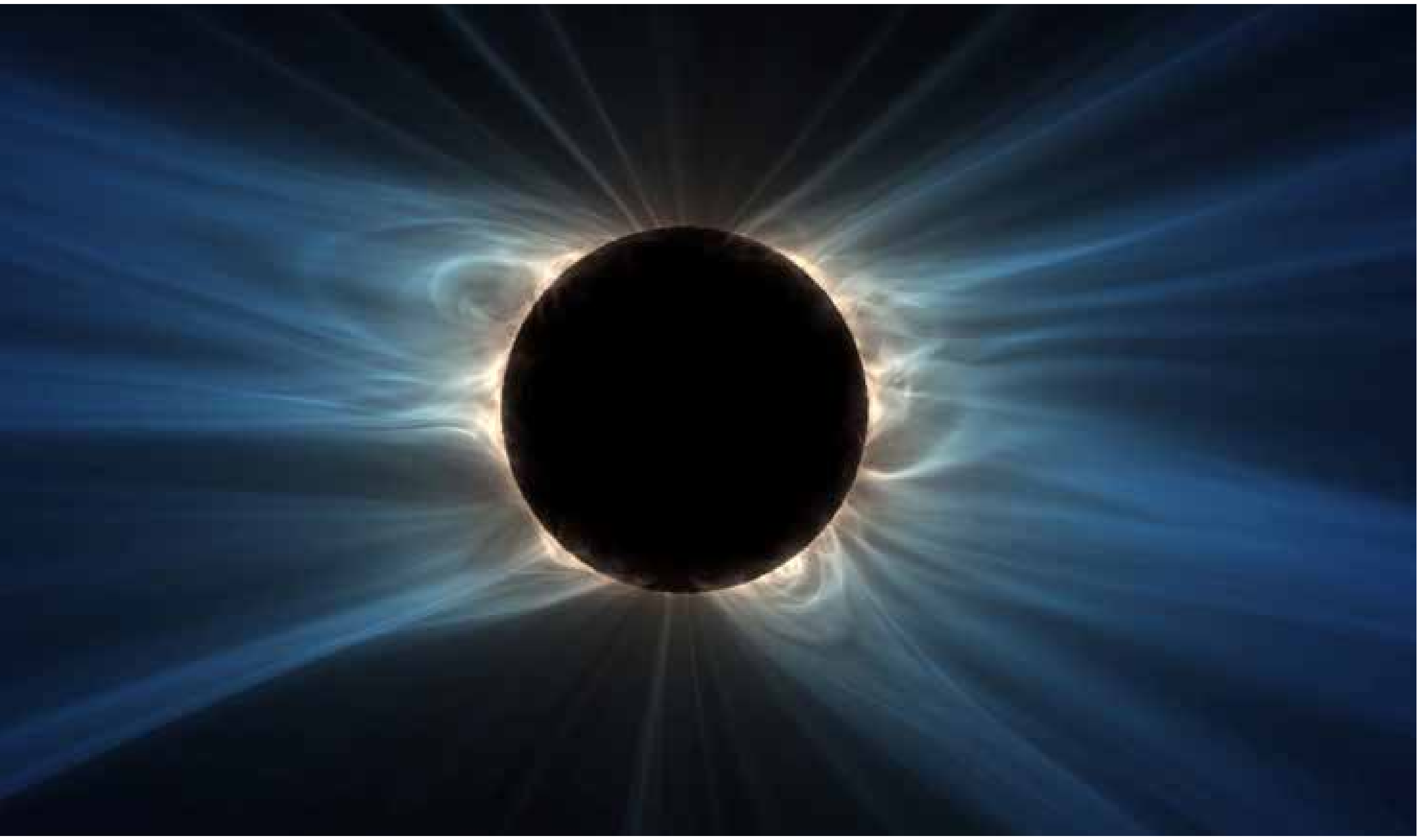
"Easy access to HPC resources is essential in advancing Earth science research," said Viswanath Nandigam, principal investigator of OpenTopography. "The availability of HPC resources at SDSC like the Expanse supercomputer will play a critical part in solving Big Data cyberinfrastructure challenges of the future."

"We calculated meter-scale topographic change at the largest spatial extent yet by using data collected by the U.S. Geological Survey's 3D Elevation Program (3DEP). This program is likely to have national coverage in a couple of years," said Scott, a co-investigator of OpenTopography. "We solved a number of Big Data geospatial challenges that we hope can be applied to this national-wide dataset to address future questions in hydrology, biomass change and hazards as well as the interaction between people and the planet's landscape."

As the USGS's 3DEP activity aims to provide topographic lidar coverage of the entire lower 48 states by 2023, there is growing potential to perform additional work evaluating large-scale topographic change. These studies will be crucial in better characterizing Earth's landscape and how it changes over time.



Indiana statewide topographic change with high resolution lidar topography collected in 2011–2013 and 2016–2020. (A) The differenced bare earth surface and (B) the natural and built surface, including the bare surface, vegetation, and structures. (The distinct north-south and east-west oriented lines do not represent change and are an error artifact from the lidar data). Credit: Chelsea Scott et.al 2022



Predictive Science Inc. Researchers Use Expanse for Sneak Peek of Extended Solar Corona

On Dec. 4, 2021, a total eclipse of the sun occurred at 07:33 (Universal Time) over Antarctica and parts of the South Pacific near the southern tip of Chile. The solar corona—visible to the naked eye only during a total eclipse—was viewable for just over one minute. Thanks to the handiwork of Predictive Science Inc. researchers, simulations were created using the Expanse supercomputer at SDSC to render a preview of what the spectacle might look like.

This prediction, posted by the researchers one week before totality, was based on a state-of-the-art computer simulation of the tenuous, magnetized outer atmosphere of the sun known as the solar corona. In addition to providing a sneak peek for eclipse chasers, such predictions aid scientists who are planning eclipse observations from the ground, sea and air. Eclipses also provide a unique opportunity for researchers to test the accuracy and predictive capability of physical models of the solar corona. Such models provide insights about how the corona is heated and how it drives the structure and dynamics of the inner-heliosphere, including Earth-affecting disturbances known as space weather.

“Expanse was an essential resource—essentially because of its unique hardware architecture and rapid turnaround for mid-scale simulations,” said Cooper Downs, an astrophysicist at Predictive Science Inc. “Although we have allocations on other supercomputers, without Expanse we would have had to start everything several days earlier to make sure all the runs and renders could be completed on time. This would mean the solar observations used to drive the model would be even more out of date, risking the accuracy of the prediction.”

Downs, whose research focuses on understanding thermodynamic and magnetic processes in the solar corona, is particularly interested in the improvement and validation of numerical models through direct comparisons to observational data. He said that Expanse allows him to run larger simulations and create more accurate diagnostics.

“Because of the large core and memory count on Expanse nodes, we have the ability to run several cases of various degrees of complexity and size—some quite large—with rapid turnaround, which was essential for getting the final prediction together,” Downs said.

About Predictive Science Inc.

Predictive Science Inc. delivers state-of-the-art scientific solutions, with research programs that focus on the development and applications of sophisticated magnetohydrodynamic models of the sun’s corona and heliosphere. With a goal of transitioning models into operational codes that can predict space weather conditions with advance warning, its programs support a number of NASA missions, including Solar TErrestrial RElations Observatory (STEREO) and Solar Dynamics Observatory (SDO), and they are performed under the auspices of NASA, the NSF and the Air Force Office of Sponsored Research (AFOSR).

A Glimpse Inside Expanse

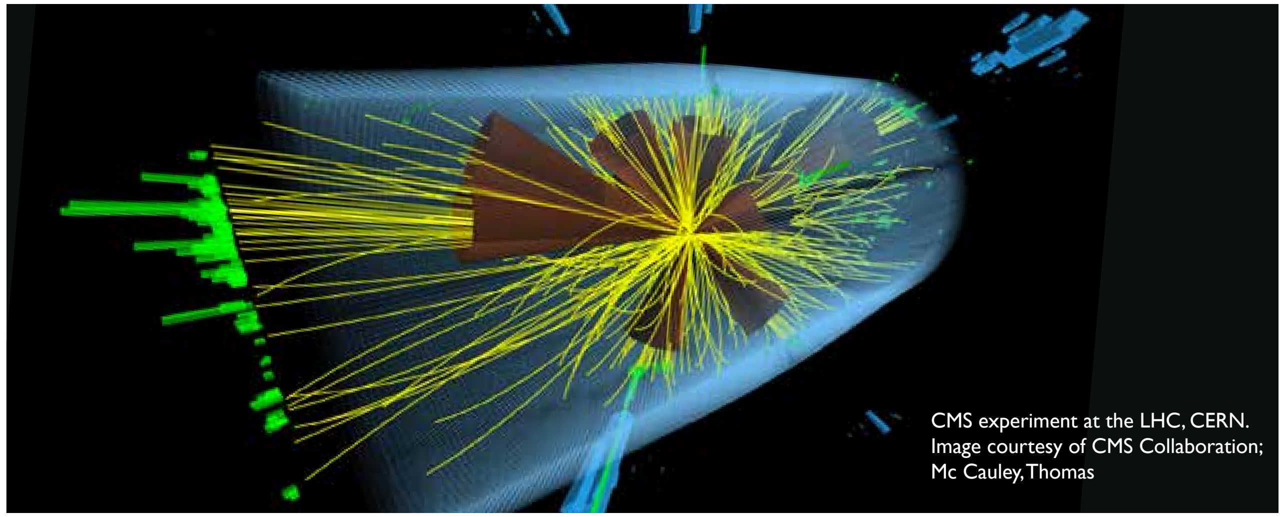
With innovations in cloud integration and features such as composable systems, as well as continued support for science gateways and distributed computing via the Open Science Grid (OSG), Expanse allows researchers to push the boundaries of computing and speed time to discovery.

A key innovation of Expanse is its ability to support composable systems—the integration of computing elements such as a combination of CPU, GPU and other resources into scientific workflows that may include data acquisition and processing, machine learning and traditional simulation. Expanse also supports integration with public cloud providers, leveraging high-speed networks to ease data movement to and from the cloud, with a familiar scheduler-based approach.

Expanse is designed for modest-scale jobs—referred to as the “long tail of science”—from one to several hundred cores, including high-throughput computing jobs via integration with the OSG, which can have tens of thousands of single-core jobs. Most disciplines—from multi-messenger astronomy, genomics and the social sciences, to more traditional ones such as earth sciences and biology—depend on these medium-scale, innovative systems for much of their productive computing.

Expanse‘s standard compute nodes are each powered by two 64-core AMD EPYC 7742 processors and contain 256 GB of DDR4 memory, while each GPU node contains four NVIDIA V100s connected via NVLINK, and dual 20-core Intel Xeon 6248 CPUs. Expanse also has four 2 TB large memory nodes. The entire system, integrated by Dell, is organized into 13 SDSC Scalable Compute Units (SSCUs), comprising 56 standard nodes and four GPU nodes, and connected with 100 GB/s HDR InfiniBand. Direct liquid cooling (DLC) to the compute nodes provides high core count processors with a cooling solution that improves system reliability and contributes to SDSC’s energy efficient data center.

Expanse is one of SDSC’s newest NSF-funded supercomputers. It supports SDSC’s theme of “Computing without Boundaries” with its data-centric architecture, public cloud integration and state-of-the-art GPUs for incorporating experimental facilities and edge computing.



CMS experiment at the LHC, CERN. Image courtesy of CMS Collaboration; Mc Cauley, Thomas

Doubling Compute Capacity for Data-Intensive Physics with Microsoft Azure

SDSC worked with the Open Science Grid (OSG) to use Microsoft Azure Cloud Services to expedite a set of high-profile data analyses in particle physics. The outcome more than doubled the compute capacity UC San Diego is providing to the Compact Muon Solenoid (CMS) collaboration.

The CMS experiment at CERN's Large Hadron Collider (LHC) is one of the largest data producers in the scientific world. Its standard data products are centrally produced and used frequently by competing teams within the collaboration, which is made up of more than 200 institutions in 40 countries. The OSG is a federated infrastructure allowing many independent resource providers to serve various independent user communities.

Teaming up SDSC and OSG to use Azure resulted in half a dozen CMS collaborators at UC San Diego, UC Santa Barbara, Boston University and Baylor University—accomplishing in a few days what would normally take them several weeks. This in turn motivated them to pursue more compute-intensive studies that would not have been possible before the CMS collaboration. Cloud integration both accelerated the science and allowed researchers to pursue science that otherwise would have been out of reach.

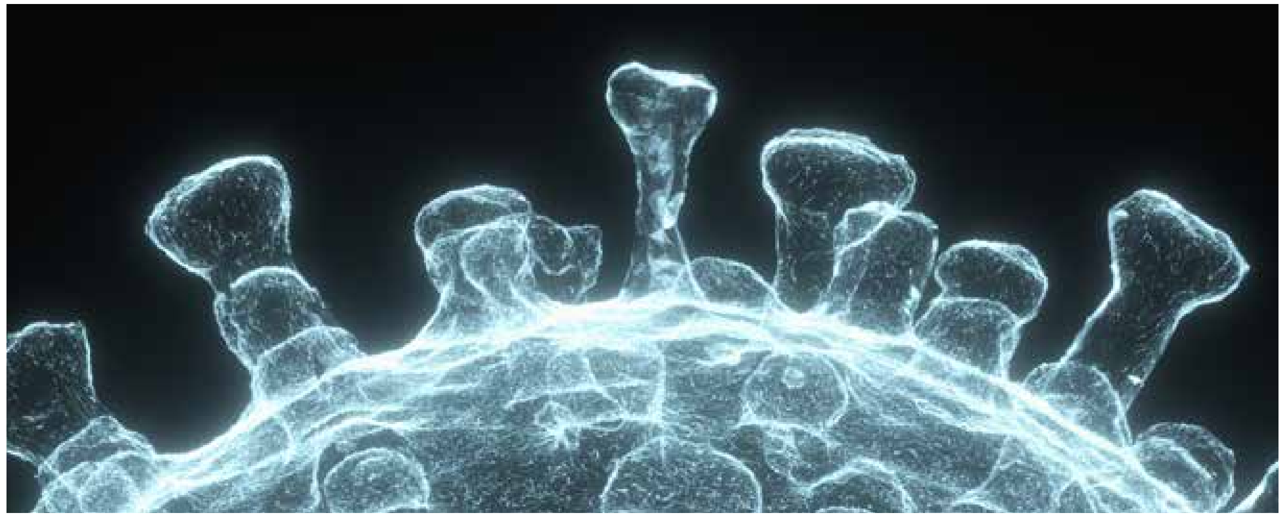
"Our collaborators already had access to many on-premises resource providers through OSG, so adding commercial clouds as resource providers was a natural evolution," explained SDSC's Igor Sfiligoi, first author of a Conference on Computer Networks, Big Data and IoT (Internet of Things) paper titled, "Data intensive physics analysis in Azure cloud," co-authored by SDSC Director Frank Würthwein and SDSC Data Scientist Diego Davila. "The OSG technology allows for easy integration of cloud compute resources, and dedicated data caching infrastructure ensures high efficiency for the data-intensive CMS compute jobs."

According to Würthwein, OSG provides the necessary trust and technical mechanisms—the glue—that allows seamless integration of the many resource providers and user communities without combinatorial issues. He explained that all resource provisioning details in OSG are abstracted behind a portal—the Compute Entrypoint (CE).

"The portal implementation used is HTCondor-CE and relies on a batch system paradigm, translating global requests into local ones. After successful authentication, the presented credential is mapped to a local system account and all further authorization and policy management is handled in the local account domain," said Würthwein. "Several backend batch systems are supported; we used HTCondor for this work, due to its proven scalability and extreme flexibility."

According to Sfiligoi, given the data-intensive nature of most CMS analyses, the remote nature of cloud resources required the deployment of content delivery network services in Azure to minimize data-access related inefficiencies. This was particularly urgent given the use of multiple cloud regions, spanning both the U.S. and European locations. The caches used performed greatly, keeping the central processing unit (CPU) utilization on par with "on-prem" resources.

This work was supported in part by the National Science Foundation (grants OAC-2030508, MPS-1148698, OAC-1826967, OAC-1836650, OAC-1541349, PHY-1624356 and CNS-1925001) and credits provided by Microsoft.



SDSC and Melax Tech Construct AI-assisted Application Using COVID-19 Data Sets

Melax Tech, along with the SDSC, announced the securing of a National Institutes of Health (NIH) Small Business Innovation Research contract to develop and enhancement of innovative user-focused informatic tools for use in basic and clinical research on infectious, immune and allergic diseases.

The Phase 1 contract, valued at $300,000, was awarded by the National Institutes of Health (NIH) to develop a COVID-19 Dataset Knowledge Graph (COVID-DKG) to facilitate integrated analysis of heterogeneous COVID-19 data sets. The work is being done in partnership with SDSC, where researchers contribute experience gained in developing Knowledge Graphs by participating in the NSF Open Knowledge Network initiative as well as other COVID-19 related projects.

Despite existing efforts making COVID-19 data sets available for research uses, the volume and the heterogeneity of COVID-19 data make it difficult for researchers to find, understand and reuse relevant data sets for their research purposes.

According to Jingcheng Du, director of Natural Language Processing (NLP) research at Melax Tech, when researchers want to conduct studies across domains, they often spend tremendous effort and time finding and learning about available data sources.

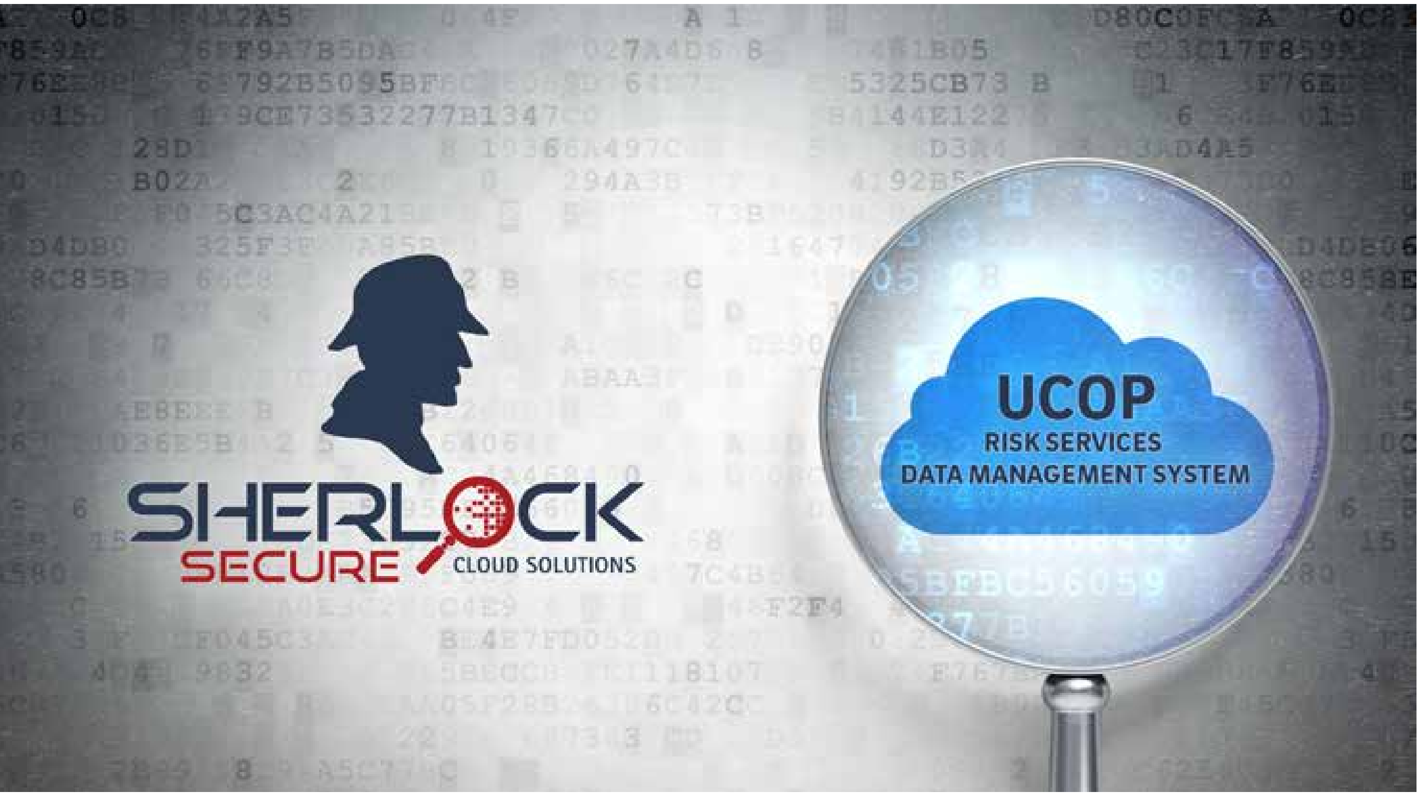
"The COVID-DKG has built-in components that support semantic search, visualization and integrated online analysis on open COVID-19 data sets, thus facilitating research and development in combating the COVID-19 pandemic," said Du.

Peter Rose, director of the Structural Bioinformatics Lab and lead for Bioinformatics and Biomedical Applications at SDSC, added, "Knowledge Graphs connect the dots by providing links among otherwise isolated COVID-19 data sets, enabling researchers to gain new insights. For example, a data set published by one journal may be joined with another data set in a different data repository to facilitate new types of analyses."

About Melax Tech

Melax Tech's award-winning flagship product CLAMP, the Clinical Language, Annotation, Modeling and Processing toolkit, was developed by a team with more than 25 years of award-winning clinical NLP experience. CLAMP enables recognition and automatic encoding of clinical information in narrative patient reports. Melax Tech was founded in 2017 and has more than 700 organizations using their technology. More information about Melax Tech's solutions is available on the company website.

Melax Tech
World leaders in biomedical natural language processing technology.
www.melaxtech.com



Sherlock Partners with UC Office of the President to Deliver Data Platform in AWS Cloud

The Sherlock Division at SDSC participated in a partnership to successfully re-architect and migrate the University of California Office of the President's (UCOP) Risk Services Data Management System (RDMS 1.0) from an on-premise, Hadoop-based platform to a serverless, data lake platform in the Amazon Web Services (AWS) Cloud (RDMS 2.0).

The 18-month production deployment of RDMS 2.0—a result of the strong and dedicated collaborative effort among SDSC's Sherlock, UCOP Risk Technology Services, UCOP Technology Delivery Services (TDS) and Kwartile—enabled the team to efficiently meet its goal and recently deliver the cloud-based RDMS 2.0.

The initial RDMS 1.0 service was conceptualized in 2015 and hosted within Sherlock's secure enclave at SDSC as a Hadoop-based data platform. As the project and its needs evolved, the natural progression of RDMS 1.0 was to refactor it to a commercial cloud to allow for the adoption and integration of new cloud-based technologies and services that would modernize the data platform, yield significant cost savings, enhance security and improve scalability. Driven by these goals, the team decided to undertake a Proof of Value (POV) effort that validated the feasibility and benefits of the technical approach while securing the necessary buy-in from the stakeholders. This was followed by a longer, more detailed project engagement to perform the full migration of the current platform to the new cloud-based solution.

"This was an excellent collaboration and a well-coordinated effort between the various teams supporting the RDMS transition from Sherlock's on-premise tenant to its AWS cloud enclave. Due to license renewal constraints, the project was completed within an accelerated timeline. This project is a win-win for the executive sponsor, resulting in both the modernization of the data platform and cost savings achieved through eliminating licensing and other operational costs." said Nilofeur Samuel, director of Risk Technology Services at UCOP.

Sherlock and its partners' overall objective for RDMS 2.0 was to create a well-architected solution that focused on delivering value to customers while addressing the following key attributes:

Modernization, Scalability, Reliability and Performance

Adopt a cloud-native, serverless data management stack that leverages the high availability and performance of AWS Cloud including:

- Data stored as objects in AWS's affordable, highly reliable and scalable data store service (S3)
- AWS Glue, a dynamic compute services that extract, transform and load (ETL) data for use by business intelligence reporting
- Use of Athena, a serverless, pay-as-you-go, interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL

Security

A defense-in-depth strategy was employed to secure data including:

- Separate environments for production and non-production data
- De-identified data in non-production environments
- Custom encryption at-rest per data environment
- Role-based, fine-grained access control to tables and columns in risk data store
- Data versioning and replication

Cost Savings

The migration from RDMS 1.0 to RDMS 2.0 is projected to save the program approximately $2 million over the next five years. These cost savings are primarily achieved by reducing licensing costs, eliminating large capital investment in physical hardware and realizing efficiencies in staffing resulting from the move from on-premise to the cloud. Specifically:

- RDMS 1.0 used proprietary licensed software, Cloudera, running on fixed infrastructure
- RDMS 2.0 runs as AWS Cloud services with a pay-as-you-go model

"As custodians of systemwide data for the university, it is incumbent upon us to continuously explore options for managing data securely, more economically and with greater flexibility and scalability. In recent years, the offerings by commercial cloud service providers, such as AWS, have become viable options for managing data that are congruent with the aforementioned tenets of our mission. Through a strong partnership between Sherlock, Kwartile and UCOP Technology Delivery Services, we were able to leverage the core competencies of each team toward a successful implementation of a modern, cloud-based and highly secure data management platform that could serve as an all-encompassing, strategic and forward-looking approach to data management," said Hooman Pejman, data architect at UCOP. "In my view, the key to our success was our collective diligence in exploring, identifying, selecting and orchestrating the appropriate services offered by AWS, based on a serverless architecture and a pay-as-you-go model."

Kwartile's data engineering solutions provided automated tools for data and metadata migration and helped update and optimize the data curation jobs to run on AWS cloud native services. "These migration tools provided a comparison report of source and target, which improved data quality and significantly reduced data validation time. Projects of this nature are complex and would not have been successful without the proper collaboration and technology expertise provided by Sherlock and UCOP teams. This was a true team effort from everyone involved, and an outcome of our long-standing partnership," said Kwartile's Krishna Katikaneni.

According to Sandeep Chandra, executive director of Sherlock Cloud at SDSC, while the individual cloud services are reliable, the real work is in the orchestration and configuration of these services which are sufficiently complex that no human could correctly and reliably maintain their state.

"Sherlock provided a platform that allowed the team to define infrastructure as code with automated deployments based on changes to a shared code base, including manual approval processes as gate keepers to control configuration changes. This assures the solution is repeatable, auditable, can be rolled back to a previous state and can easily adapt to frequent incremental change," Chandra explained. "This re-usable infrastructure as code paradigm allows Sherlock to use the same building blocks and processes adapted and customized to the specific needs of different projects across various engagements."

Sherlock Cloud Solutions and Services
https://sherlock.sdsc.edu/

CICORE Adds New Strategic Partnerships Position

In late 2021, the Cyberinfrastructure and Convergence Research and Education Division (CICORE) at SDSC added a new position—Director of Strategic Partnerships. The role is filled by Melissa Floca, whose job is to build cross-institutional and cross-sectoral alliances in support of societal-scale technological innovation.

Floca is no stranger to an inclusive approach to the process of community-based problem-solving. She was named the 2021 International Leader of the Year by the San Diego Regional Chamber of Commerce for her previous work at the UC San Diego Center for U.S.-Mexican Studies and Kroc Institute for Peace and Justice at the University of San Diego. There, she led initiatives focused on innovation and inclusion at the U.S.-Mexico border. Through her work on regional border issues, she has collaborated with a number of researchers at SDSC, including the WIFIRE team, led by Ilkay Altintas.

"We are very excited to have Melissa ramp up our use-inspired research efforts involving diverse communities," said CICORE Division Director Ilkay Altintas.

Floca's responsibilities in this new position center around what the NSF identified in 2016 as one of the 10 Big Ideas for its future investments—convergence research. This type of research is driven by a specific and compelling problem solved through a process that includes integration across disciplines for the good of society.

"Convergence is the focus of this particular role and we want to be intentional about integrating innovative and sustainable solutions into society," said Floca. "The foundation for all of the work that I am doing is the idea of broadening participation and bringing underrepresented communities and individuals into our work, because any technological solutions that we build are only going to be as robust as the diversity of our collaborators."

In her inaugural role, Floca is building partnerships with a focus on meeting societal challenges through translation of data and cyberinfrastructure into practice. Her approach is organized around four pillars:

- broadening participation of diverse individuals and organizations in defining challenges and developing solutions;
- bringing disciplines together to innovate;
- working with partners who can integrate solutions into existing systems across society and
- developing collaborations to create business models that are sustainable at scale.

"To be entirely laser-focused on partnerships for impact is really exciting to me," said Floca.

Floca cited WIFIRE as a good example of convergence research that brings together partners from the fire management and fire science communities, as well as data scientists and the artificial intelligence community, to tackle the challenges posed by megafires that seasonally threaten the nation.

She also referenced her previous work on the challenges that face border communities in San Diego, Tijuana and beyond.

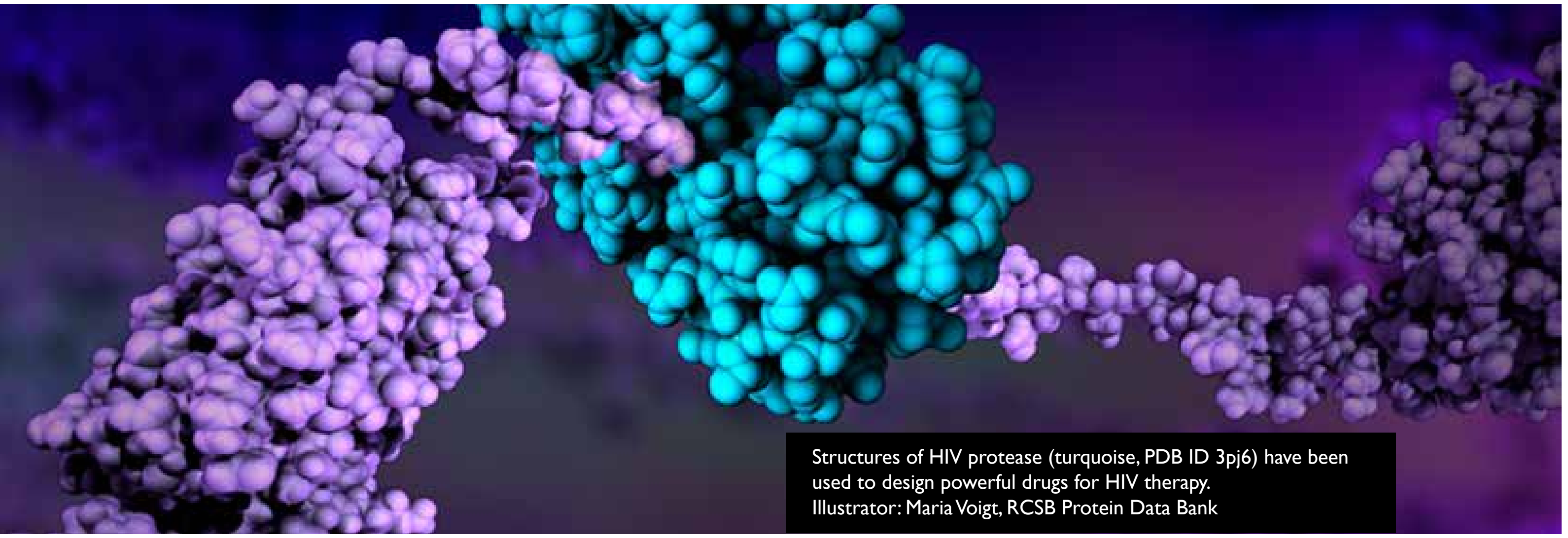
"Our location at the U.S.-Mexico border means that in our region we experience global challenges locally in a way that most places do not, from water resource-related challenges to supply chain disruptions or COVID-19. The border is a very rich learning laboratory for data-driven research, as well as data science-focused education and workforce development," Floca said.

According to Altintas, SDSC has a wide range of expertise that can be applied to solve big societal challenges and train students in experiential settings involving real problems. "Melissa brings a unique experience at a time when our team is committed to scaling our convergence research activities," she said.

Floca noted that there are increasingly urgent and complex societal challenges with few workable solutions and SDSC can play a pivotal role as a leader in scalable computing, artificial intelligence and cyberinfrastructure.

"I'm looking forward to working alongside our PIs to build partnerships for impact at the societal scale. I can't think of any more important way to use my time," said Floca.

Melissa Floca at a presentation to the California Senate Select Committee on California-Mexico Cooperation related to her work at the Center for U.S.-Mexican Studies. Image courtesy of Melissa Floca.



Structures of HIV protease (turquoise, PDB ID 3pj6) have been used to design powerful drugs for HIV therapy. Illustrator: Maria Voigt, RCSB Protein Data Bank

Protein Data Bank Worldwide Collaboratory Includes New Tech Tool for Researchers in Asia

Established in 1971 as the first open access digital data resource for biology and medicine, the Protein Data Bank (PDB) is a leading global resource for experimental data integral to scientific discovery. The PDB was founded by Board of Governors Distinguished Professor Emerita of Chemistry and Chemical Biology Helen Berman at Rutgers-New Brunswick. Berman also established the Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB), which operates the U.S. data center for the global PDB archive, and makes PDB data available at no charge to all data consumers without usage limitations.

In 1998, the RCSB PDB moved the PDB to UC San Diego—specifically to SDSC. Jose Duarte, who manages the PDB site at SDSC, reflected on the ways this worldwide collaboration has transformed how scientists collect and share their structural biology data.

"Not only does the PDB continue to showcase the power of open data and community...but thanks to open data and standards defined by the PDB, we have also witnessed the birth of a thriving sub-branch of bioinformatics known as structural bioinformatics, which is a direct consequence of the existence of the PDB."

PDB's evolution continues as announced during the SupercomputingAsia 2023 conference, when SDSC and the Singapore Advanced Research and Education Network (SingAREN) signed a Memorandum of Understanding to work toward deploying in the sovereign island/city-state a data cache – a data block for storing information for easy re-access. SDSC will contribute a high-performance server that will be hosted at the SingAREN Open Exchange, located at Equinix SG3. SingAREN will provide the high-speed international connectivity for the server in the region.

Deploying the cache server at SingAREN will provide researchers in Asia – particularly those working in the fields of genomics, climate science and materials science – with faster and more efficient access to data. This in turn will enable quicker and more efficient cutting-edge research and discoveries. Over the next three years, the two organizations will actively work together in support of the Open Science Data Federation and the RCSB PDB.

Researchers ranging from computational chemists to artists have utilized PDB for their work. For example, Rommie Amaro, distinguished professor of theoretical and computational chemistry at UC San Diego, has shared insights into the molecular piece parts of the SARS-CoV-2 virus, and used the data, together with molecular dynamics simulations, integrative modeling and AI to understand how the viral spike protein opens.

"Importantly, the PDB showed the biological world how to think about, organize, collect and develop data into a useful ecosystem—this ecosystem shows the centrality of PDB data across different biological domains and scales," Amaro said.

Another "power" user of PDB is Artist/Scientist David Goodsell, a professor of computational biology at the Scripps Research Institute and a research professor at Rutgers University. "The PDB is an essential resource for education and outreach, providing a detailed look at the molecules that perform the processes of life," he said.

PDB's Impact

More than $5 billion in funding has been provided by the National Institutes of Health (NIH) to structural biologists in the U.S. who have generated more than 50,000 of the structures currently available from the PDB.

Biomedical researchers using the structure data stored in the PDB have published more than two million scientific papers, some of which have helped researchers and pharmaceutical companies tackle major health challenges, including heart disease, cancer, diabetes, Alzheimer's disease and HIV-AIDS.

Ann Stock, distinguished professor in the Department of Biochemistry and Molecular Biology at Robert Wood Johnson Medical School and associate director of the Center for Advanced Biotechnology and Medicine (CABM), said the data shared through the PDB is central to understanding biological systems at the molecular level – an integral part of drug development being done to treat human diseases by both biotechnology and pharmaceutical companies.

"While some investigators wanted to keep information to themselves to guide their own investigations in the early days of structural biology, the PDB enabled data sharing and had support of the government and the academic scientific community, who understood that this information was critical to researchers throughout the world," Stock said.

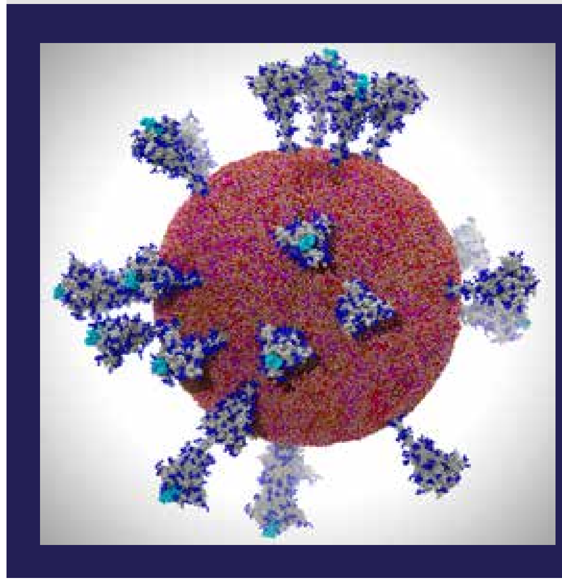
Today, this means that structural biology researchers who want to publish in peer-reviewed scientific journals must share their data via the PDB.

"Sharing of scientific data is something that has evolved with a lot of progress over the last couple of decades," Stock said. "The PDB was one of the first databases that provided a comprehensive set of data for a particular field and set policies early on about what needed to be shared."

According to Stephen Burley, professor and Henry Rutgers Chair at Rutgers-New Brunswick and director of the RCSB PDB and the Rutgers Institute for Quantitative Biomedicine, open access to 3D structure information from the PDB facilitated discovery and development of more than 90 percent of the 210 newly approved by the U.S. Food and Drug Administration (FDA) between 2010 and 2016.

"Looking more closely at the 54 new anti-cancer drugs approved by the FDA in 2010 to 2018, revealed that more than 70 percent of them were the products of structured-guided drug discovery accelerated by open access to PDB structures of the drug targets."

The PDB is managed by the Worldwide Protein Data Bank partnership, with data centers in the U.S., Europe and Asia.



UC San Diego's Distinguished Professor in Theoretical and Computational Chemistry Rommie Amaro and her team combined high-resolution PDB data with lower-resolution cryo-electron tomography data in order to model the SARS-CoV-2 virion in atomic detail. Amaro explained that she and others have used PDB data to help guide computer-aided drug design methods, which are helping researchers more efficiently and effectively develop new and safer medications. "Put another way, the PDB is so central to biological and biochemical research, it's nearly impossible to imagine a world without it," she said. Credit: Rommie Amaro, UC San Diego

Data Science Students Develop Simulation System for Keeping Schools COVID-safe

For the first two years of the pandemic, a group of UC San Diego researchers met weekly with epidemiologists at the County of San Diego Health and Human Services Agency (HHSA) to discuss COVID-19 dynamics, analyze populations at higher risk and explore the county's pandemic response and new ways to mitigate the infection. A collaboration formed at the start of the pandemic and led by SDSC researcher Ilya Zaslavsky and a team of UC San Diego undergraduate data science students, resulted in an agent-based simulation system to assist in COVID-safe school re-openings within San Diego County.

While a variety of high-level policies were considered to make school re-openings as safe as possible, little was known about potential infection spread in a school setting and the efficacy of mitigation measures. At the same time, both teachers and parents were anxious to learn what might happen with children in specific schools, given each site's variation in design, resources and mitigation plans.

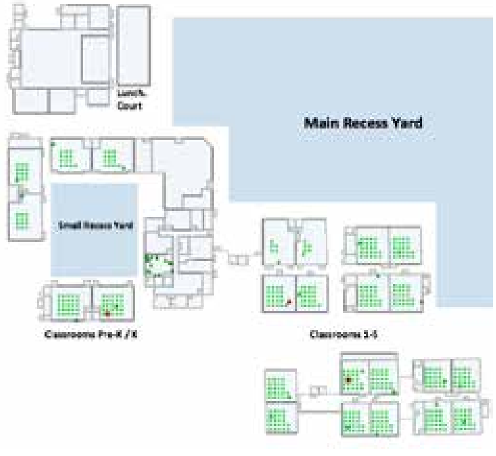
Using an early online version of a site simulator, school officials simulated interactions between agents—students and teachers in different grades—as they participated in different types of activities throughout 15 school days: learning at individual desks during class time, group activities, recess time and having lunch in the school cafeteria or classroom. To assess infection risks due to aerosol and droplet transmission, the model used real school floor plans, layouts and capacities of classrooms, cafeterias and recess areas.

Schools were mostly closed for on-campus instruction at the time of these initial simulations, so the modeling team had to rely on data from other countries, from literature and from other fluid dynamics models to generate these initial simulations.

“The simulations allowed us to pinpoint areas in schools that would present higher COVID-19 transmission risks, and to evaluate relative importance of non-pharmaceutical interventions such as wearing masks, reducing class sizes or canceling lunch in the cafeteria and moving it to classrooms,” said Zaslavsky, who is director of the SDSC Spatial Information Systems Laboratory.

According to Zaslavsky, the spatially explicit, agent-based modeling of COVID-19 transmission at schools allowed individual sites and districts to test their plans and match them with their specific spaces, resources and population.

Within several days after the team demonstrated the model in September 2020 to nearly 400 San Diego educators, on a call organized by the County of San Diego, the simulations were run over 300 times. Leslie Ray, senior epidemiologist with the HHSA, said, “At a time when schools were struggling with how to reopen safely, the SDSC team developed an easy-to-use, free tool that allowed schools to test their plans on pixels rather than pupils.”



Schematic representing the layout and floor-plan of the school for model runs.

From the time of the first simulations, the core modeling team, which included UC San Diego data science majors Kaushik Ganapathy, Bailey Man, Johnny Lei and Eric Yu, continued to enhance the model based on the growing literature on COVID-19 epidemiology. Over the year, the team added students and used feedback and advice from the County of San Diego HHSA and from Howard Taras, MD (UC San Diego Pediatrics), key advisor to San Diego schools on managing COVID-19 infection.

From classrooms to school buses

A key concern of school administrators was evaluating risks of COVID-19 transmission on school buses. To extend the classroom model, the transportation division of the San Diego Unified School District (SDUSD) hosted a “data science field trip” to let the data science students take measurements of different types of school buses used in San Diego. They also ran a smoke machine in a moving bus to record air flows at different speeds. Simulating transmission on school buses at the request of SDUSD, the team explored how different seating patterns, bus occupancy, mask wearing and ventilation with windows and hatches open or closed, would affect infection risks.

“In school buses where students with special health care needs cannot always keep their masks on, these models have been very helpful to determine where students should be seated to maximize safety,” said Dr. Taras.

At the same time, using the National Science Foundation's Extreme Science and Engineering Discovery Environment (XSEDE) with help from the Science Gateways Community Institute, the model components were installed on SDSC's Comet and Expanse supercomputers and accessed via an Apache Airavata gateway, which allowed the researchers to compute model scenarios much faster and with more input parameters, such as counts of students and staff in each cohort, testing regimens and vaccination rates for students and teachers, mask types and adoption, class occupancy and ventilation characteristics. The system is accessible via the Geographically assisted Agent-based model for COVID-19 Transmission, or GeoACT (supported by the National Science Foundation, award no. 2139740).

Making e-decisions

A recent outcome of this collaboration was the e-Decision Tree that is now used by schools around San Diego County to generate instructions to students or school staff who showed COVID-19 symptoms or have been in close contact with infected individuals. The e-Decision Tree was developed by Zaslavsky working with UC San Diego undergraduate data science student Alice Lu. It follows guidance from CDC, the California Department of Public Health and San Diego County public health orders, and provides schools with an easy-to-use tool to decide when a student or staff should come back to school after isolation or quarantine, and on which dates to get tested. It is not only linked from the San Diego County Office of Education (SDCOE) COVID-19 website, but is also a valued resource on the California Safe Schools for All website.

“By answering a series of five or six brief questions generated by this electronic tool, SDSC has allowed school nurses and primary care pediatricians to replace a time-consuming and complex set of public health regulations with a 30-second survey. And this amazing facilitative device repeats this gift of time several times a day, child after child,” said Dr. Taras.

Triton Shared Computing Cluster Gets Tuned-up for Users

First launched in 2013, the Triton Shared Computing Cluster (TSCC) has been a critical resource for the community of UC San Diego researchers—as well as UC Riverside and UC Merced teams—providing high-performance computing services to support modeling, simulation and data analytics.

While many computing clusters consist of a set of individual computer servers or "nodes" that are networked together and have software installed that allows them to operate in concert for parallel computing, TSCC operates on the "condo cluster" model—researchers purchase nodes and contribute them to the cluster, forming a community-owned and shared resource.

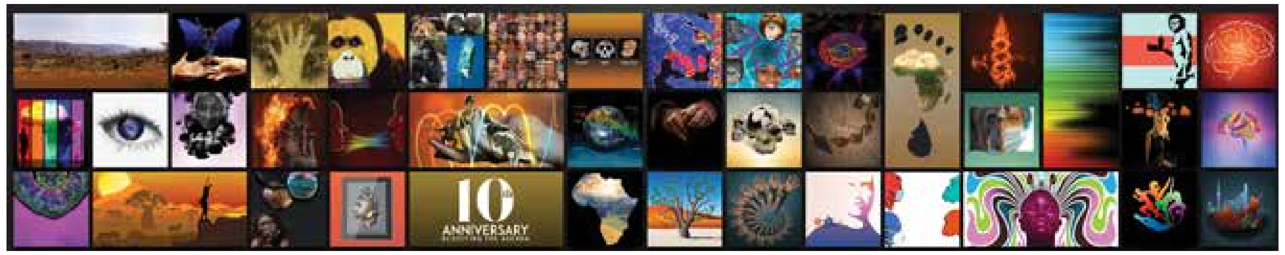
After almost a decade of operation, the TSCC networking system has been upgraded and reconfigured to better serve users. A new set of standard node offerings has also been defined to better control heterogeneity, older nodes are being decommissioned and replacement nodes are being installed. A plan was also developed to replace the high-performance parallel and "home" file systems. The new TSCC system offers higher performance and more energy efficiency.

The way TSCC has worked is that researchers purchase one or more nodes to gain access to the entire cluster, which in most cases is a much larger resource than they could purchase or justify just for their group. A nominal monthly fee (per node) affords access to shared infrastructure on TSCC, including high-performance storage, as well as professional system administration and user support. The number of TSCC nodes has been as high as 300, serving over 30 research groups and hundreds of faculty members, scientists and students across most of the academic departments at UC San Diego.

In addition to the condo cluster, TSCC has a separate "hotel" section—a smaller portion of the cluster that is available to researchers through ad hoc recharges of computing time. The hotel section serves researchers who have short-term or "bursty" computing needs, or who do not have funds available to participate in the condo program.

"The team was very busy with day-to-day operations, but in order to maintain the high level of performance and reliability our users had come to expect, we knew we had to find some time and plan for the future," said TSCC Lead Subhashini Sivagnanam, referring to the upgrade and reconfiguration.

While it began years ago as a relatively modest effort, TSCC has grown into a "mission critical" resource for UC San Diego researchers who utilize high-performance computing. With completion of the planned upgrades, TSCC will be well positioned to provide high-performance scientific and technical computing support in the years to come—giving researchers the tools they need to continue advancing the frontiers of science.



CARTA Achieves Rare Milestone via Key Partnerships

A symposium series on human origins, created by the Center for Academic Research and Training in Anthropogeny (CARTA), recently reached a rare milestone of more than 40 million online views of its recorded sessions. This milestone, which ranks CARTA as University of California Television (UCTV)'s most popular science series and second most popular series overall, was accomplished by a team from UC San Diego, SDSC, the Salk Institute for Biological Studies (Salk) and UCTV/UCSD-TV.

CARTA offers free public symposia that feature multidisciplinary experts from around the world discussing topics related to human origins and uniqueness. It was originally established as the UC San Diego Project for Explaining the Origin of Humans (POH) in the 1990s. Since 2008, CARTA has operated as an Organized Research Unit (ORU) and a collaboration among faculty at UC San Diego and Salk to promote transdisciplinary research investigating the origin of humans, or anthropogeny, drawing on methods from a number of traditional disciplines spanning the social, biomedical, biological, computational and engineering, physical and chemical sciences, and the humanities. Through its symposia, graduate and undergraduate education, and research collaborations, CARTA explores such topics as bipedalism, stone tool technologies, diet, human development, molecular biology, evolutionary medicine and anthropogenic climate change.

SDSC joined the POH effort in 2001 to provide the group with informatics support, and it became an official partner in 2008 when CARTA was formed.

"CARTA's cyberinfrastructure has expanded along with its rapidly growing global community to include custom web portals, public cloud services, and scientific data management supporting a variety of formats such as tomography, radiograph and curated documents," explained SDSC Cyberinfrastructure Specialist Kate Kaya, technical lead for CARTA. "I knew CARTA would be a fun project to work on when I watched CT bone scans for the Museum of Primatology on my first day. Where else could I build web-based research platforms, edit fascinating videos, see students visualize articulated skeletons, and listen in as renowned academics from around the world passionately discuss what makes us human?"

CARTA's in-person events, primarily hosted at Salk, were well attended prior to the global COVID-19 pandemic, with an average of 400 in-person attendees and 200 more viewing via live stream. UCSD-TV produced and widely broadcast these in-person symposia via UCTV and online channels. Thanks to the live stream infrastructure that SDSC and Salk established in 2012, and the high-quality UCSD-TV video production process, CARTA was well positioned to transition to online-only events at the start of the pandemic in early 2020. SDSC led CARTA's initial efforts to offer completely virtual events with live interactive expert panel discussions, working with the team at UCSD-TV to ensure the continuity of CARTA's core mission to explore and explain human origins. These online symposia have been extremely successful, with CARTA's most recent virtual event reaching roughly 660 live stream viewers from across 40 countries and territories. CARTA's live symposia are also recorded and made freely available to the public on multiple websites, including CARTA, UCSD-TV, iTunes and YouTube. These recordings have proved to be hugely popular with over 40 million views online and counting.

"CARTA's long-term partnership with the San Diego Supercomputer Center has been extremely valuable," said Dr. Ajit Varki, CARTA's founding co-director and distinguished Professor of Medicine and Cellular & Molecular Medicine at UC San Diego. "Having access to state-of-the art cyberinfrastructure and information technology expertise has played a key role in advancing CARTA's mission by allowing us to rapidly meet evolving community needs."

"The type of transdisciplinary research promoted by CARTA is the epitome of convergence research that SDSC excels in facilitating. We are privileged to be able to support the breathtaking diversity of researchers that CARTA brings together to address fundamental questions about the human phenomenon," said Chaitan Baruof the NSF.

SDSC's Hans-Werner Braun Inducted into Internet Hall of Fame

Twenty-one pioneering individuals who fundamentally changed the world by building and developing the global Internet were inducted into the Internet Hall of Fame late in 2021. Hans-Werner Braun, a research scientist at SDSC was among the engineers, physicists, mathematicians, academics and others from 11 nations recognized for their outstanding contributions to the Internet's global growth.

These global web pioneers invented the technologies that launched the Internet, expanding its reach in their own regions and worldwide, and making it more secure, reliable and accessible for millions. The Internet they helped create brought the new cohort together in a virtual induction ceremony that took place last week, where they logged on from worldwide locations to share the honor with their colleagues, about whom Internet Society President Andrew Sullivan noted: "Their contributions made it possible for us to look forward to our future, inextricably tied to the open, globally connected, secure and trustworthy Internet, and its ability to connect us reliably and consistently."

Braun was specifically inducted for his role in the design, development and operation of the National Science Foundation Network, NSFNET, and the network's subsequent growth in speed, coverage and reliability, which served as a model for Internet networks around the world and paved the way for large-scale routing.

"I appreciate being included in the Internet Hall of Fame, however, there are many others who played key roles as well who seem to be all but forgotten by now," Braun said. "I cannot even start to inclusively credit people for working with me or helping me, as there were literally at least hundreds."

After working for five years on a regional university computer network at the University of Cologne in West Germany, in 1983 Braun joined the Michigan Educational Research Triad (MERIT) team at the University of Michigan, where he played a critical role in the development of the original interim NSFNET backbone, which ran at 56 kilobits per second. He did this by installing Dave Mills' Fuzzball software to get the backbone operational, while being connected to a node via the National Center for Atmospheric Research's University Satellite Network (USAN) project, utilizing a geostationary satellite. He continued to de-facto run the backbone.

Then, in November 1987, NSF awarded the "Management and Operation of the NSFNET Backbone Network" cooperative agreement to MERIT with its joint study partners MCI and IBM as well as additional funding by the State of Michigan. The purpose of this award was to replace the original interim NSFNET backbone with a T1 (1.544 megabits-per-second) environment, almost 28 times faster than the interim network, in order to accommodate growing traffic. Another objective was more comprehensive network management.

Under that award, Braun became the co-principal investigator for the NSFNET backbone, which gave him significant oversight for the new network's design and operation. NSFNET was originally created as an academic research network, connecting five academic supercomputer centers, which included the San Diego Supercomputer Center, plus NCAR. The new T1 award was to add regional research and education networks which would then connect individual campuses, with the NSFNET backbone literally becoming the Internet backbone.

In order to achieve this outcome, Braun and others coordinated the work of MERIT, NSF, the State of Michigan, IBM and MCI, while targeting a deadline of July 1, 1988, for the T1 launch. NSF's Steve Wolff contacted Braun on June 30 to determine when the network would be ready, to which Braun responded, "The 30th is not yet over." The upgraded system was ultimately launched later that day, at around 8 p.m., when Braun himself emailed users, "The NSFNET Backbone has reached a state where we would like to more officially let operational traffic on." According to NSF, this understated message essentially "announced the birth of the modern Internet."

Braun was instrumental in implementing further speed upgrades to NSFNET, up to an initial T3 (45Mbps) prototype between Ann Arbor and San Diego in December 1990. When it became clear NSFNET would be phased out, he helped NSF devise the follow-on architecture to support academic users, then working at the San Diego Supercomputer Center.

As the co-author of seven RFCs in the late 1980s and 1990s, he had significant influence on the development of the Internet.

Always the prescient thinker, Braun created large network measurement and analysis systems in the 1990s, and starting in 2000, he pioneered remote unattended wireless links to scientific instruments as part of the High-Performance Wireless Research and Education Network (HPWREN), ranging from individual sensors in the desert to mountain-top astronomy observatories, establishing an early "Internet of Things" well before the term existed.

Now, more than 20 years later, HPWREN is still evolving, having been connected not only to researchers and educators, but also firefighters and other public safety officials.

This High-Performance Wireless Research and Education Network (HPWREN) project photo from several years ago shows Hans-Werner Braun and a colleague conducting a (then) high-speed wireless communications test in the Anza Borrego desert. Photo courtesy of HPWREN

Students showcased their finished projects to firefighters, policy makers and community leaders at the Mindshifts on Megafire Design Challenge Expo. Credit: SDSC External Relations

Design Challenge

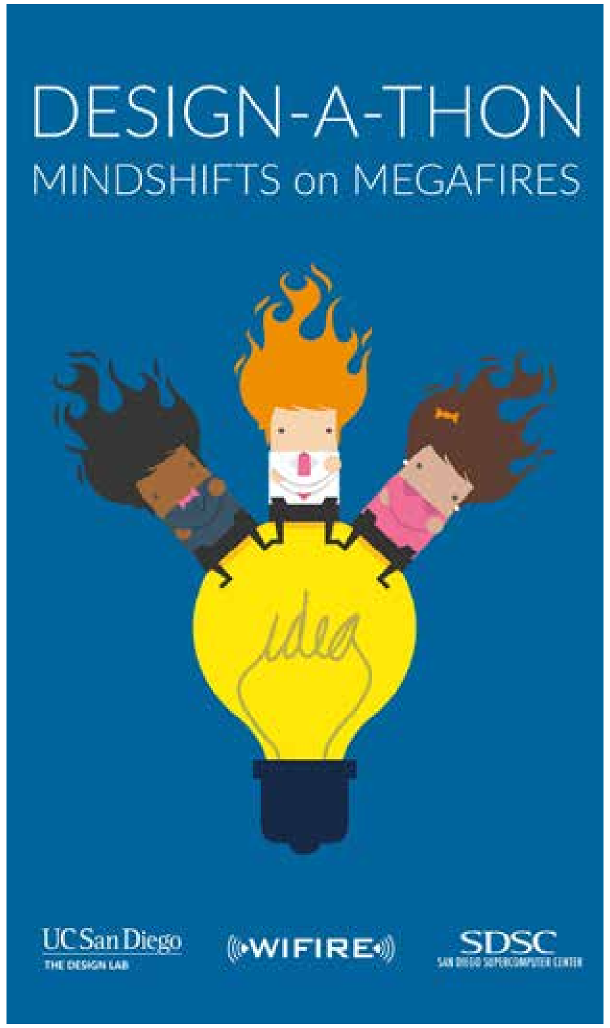
San Diego Supercomputer Center and the Design Lab at UC San Diego co-hosted the Mindshifts on Megafire Design Challenge Expo, which showcased the work of seven student teams who created concept designs to increase public understanding and acceptance of prescribed burns. Specifically, the October 2022 expo was the grand finale to share the students' efforts with a group of WIFIRE Lab stakeholders from the fire management and research community.

"Prescribed burns – the controlled use of fire under specified weather conditions to benefit ecosystems – have proven to be an important fuel treatment approach because they reduce the future risk of uncontrollable and highly destructive wildfires by reducing dangerous fuel loads," explained Ilkay Altıntaş, chief data science officer and the founding director of the WIFIRE Lab at SDSC. "In 2020, thousands of firefighters risked their lives to fight wildfires that swept across 10 million acres in the western U.S., killing dozens of people, destroying 10,000 structures and causing $15 billion in property damage. Activities like the Mindshifts on Megafire Design Challenge and Expo allow students to actively participate in helping us educate multiple communities about the importance of prescribed burns."

Prior to the expo, the teams participated in the April 2022 Design Thinking Workshop at UC San Diego led by the Design Lab. The workshop allowed 90 students to form 23 teams and create prototypes for an installation; next, they submitted a poster and video describing their concepts. The ideas ranged from virtual reality experiences to billboards at bus stops, and seven teams were selected to participate in an internship with SDSC to build their proposals into functioning prototypes. Each team submitted their ideas for evaluation to generate quantitative and qualitative feedback. The teams displayed posters and prototypes to subject matter experts during the Design@Large session on Climate Risk Reduction and Technology.

Seven teams were selected as finalists and moved on to a summer internship, during which they created working prototypes of their installations.

The interns were joined by four students from the ENLACE summer research program, which aims to encourage the participation of high school students, university students and researchers/teachers in the sciences and engineering, while promoting cross-border friendships between Latin America and the United States. The prototypes created by the student teams were on display during the expo.


DESIGN-A-THON
MINDSHIFTS on MEGAFIRES
idea
UC San Diego THE DESIGN LAB
WIFIRE
SDSC SAN DIEGO SUPERCOMPUTER CENTER

"We were amazed by the creativity and productivity of all the teams that participated in the design-a-thon and have greatly enjoyed continuing to work with teams to build their ideas into functional prototypes ranging from board games to virtual reality experiences," said Director of Strategic Partnerships at SDSC's CICORE Division Melissa Floca. "Our ultimate goal was to increase public acceptance of prescribed burns as an important tool for ending devastating megafires, and I think we achieved that."

Each team submitted their ideas for evaluation to generate quantitative and qualitative feedback. Teams displayed posters and prototypes to subject matter experts during the Design@Large session on Climate Risk Reduction and Technology.

According to Altıntaş, "We were amazed by the creativity and productivity of all the teams that participated in the design-a-thon and have greatly enjoyed continuing to work with teams to build their ideas into functional prototypes."

Supporting the Next Generation

SDSC's education and outreach programs serve students from middle school through graduate school. Additionally, SDSC offers high-performance computing online courses that attract participants from all over the world. Below is a listing of the various programs aimed at supporting aspiring science and technology leaders of the future, as well as current professionals.

ENLACE

The ENLACE summer research program at UC San Diego aims to encourage the participation of high school students, university students and researchers/teachers in research in the sciences and engineering, while promoting cross-border friendships between Latin America and the United States. Members of SDSC's CICORE Division have been involved with the program.

FORMIDABLE

An offshoot of UC San Diego's Anita Borg Leadership and Engagement (ABLE) program, this eight-week program introduces middle school students from six pilot schools to STEM careers through hands-on workshops, invited speakers, tutorials and robotics demonstrations.

PI WARS

SDSC participated in Pi Wars—an international, challenge-based robotics competition in which teams build Raspberry Pi-controlled robots and then compete in non-destructive autonomous and remote-controlled challenges. The competition encompassed several teams of middle and high school students. Because the program was virtual, to keep the students engaged the SDSC Education team led several online talks, demonstrations and virtual visits with undergraduate robotics teams.

RESEARCH EXPERIENCE FOR HIGH SCHOOL STUDENTS

SDSC's Research Experience for High School Students (REHS) program, which celebrated its 13th year in 2022, was developed to help increase awareness of computational science, science writing and related fields of research among students in the greater San Diego area. The eight-week program pairs SDSC mentors with high school students to help them obtain practical experience, while gaining exposure to career options and work-readiness skills. Capping off 2022's program was a virtual "Project Showcase," where students shared their research projects with peers, mentors, family and friends. To date, more than 525 students have participated in SDSC's REHS program.

MENTOR ASSISTANCE PROGRAM

While the REHS program takes place during the summer months, high school students interested in pursuing a career in scientific research are also invited to apply to UC San Diego's Mentor Assistance Program (MAP), a campus-wide initiative that encompasses working with experts from a vast array of disciplines. Launched five years ago by SDSC and the UC San Diego School of Medicine, MAP's mission is to provide a pathway for students to gain access to UC San Diego faculty, postdoctoral fellows, doctoral candidates and staff to mentor them in their specific fields of interest. Mentors are recruited from across campus including areas of biology, chemistry, aerospace engineering, network architectures, pharmaceutical sciences, physics, social sciences and more.

HPC@MSI

SDSC recently announced the creation of HPC@MSI, a program aimed at facilitating the use of high-performance computing (HPC) by Minority Serving Institutions (MSI). The HPC@MSI program is designed to broaden the base of researchers and educators who use advanced computing by providing an easy on-ramp to cyberinfrastructure that complements what is available at their campuses. Additional goals of the program are to seed promising computational research, facilitate collaborations between SDSC and MSIs, and to help MSI researchers be successful when pursuing larger allocation requests through the new Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, the successor to the National Science Foundation's Extreme Science and Engineering Discovery Environment (XSEDE) program.

ONLINE DATA SCIENCE AND BIG DATA COURSES

UC San Diego offers a four-part Data Science series via edX's MicroMasters® program with instructors from the Jacobs School of Engineering's Computer Science and Engineering Department and SDSC. In partnership with Coursera, SDSC created a series of MOOCs (massive open online courses) as part of a Big Data Specialization that has proven to be one of Coursera's most popular data course series. Consisting of five courses and a final capstone project, this specialization provides valuable insight into the tools and systems used by big data scientists and engineers. In the final capstone project, students apply their acquired skills to a real-world big data problem. To date, the courses have reached more than one million students around the world—from Uruguay to Bangladesh.

HPC TRAINING WEBINARS AND WORKSHOPS

A vast array of HPC training opportunities are offered each year at SDSC. In 2021-2022, more than 50 events were held that focused on familiarizing researchers with HPC systems such as Expanse, Voyager and the Triton Shared Computing Cluster. The Center's HPC programs were attended by thousands of participants, and workshop topics included running parallel jobs on HPC systems, GPU computing, parallel computing with Python, Python for data scientists, machine learning, parallel visualization, using Singularity containers for HPC and using Jupyter Notebooks for HPC and data science.

SDSC SUMMER INSTITUTE

Aimed at researchers in both academia and industry, the week-long workshop focuses on a broad spectrum of introductory-to-intermediate topics in HPC and data science. The 2022 Summer Institute gave attendees an opportunity to accelerate their learning process through hands-on tutorial classes using the Expanse supercomputer. The 2022 Summer Institute also continued SDSC's strategy of bringing advanced cyberinfrastructure to the "long tail of science" and provided resources to a larger number of modest-sized computational research projects that advance, in aggregate, a tremendous amount of scientific progress.

CYBERINFRASTRUCTURE-ENABLED MACHINE LEARNING (CIML) SUMMER INSTITUTE

The 2022 CIML Summer Institute introduced machine learning (ML) researchers, developers and educators to the techniques and methods needed to migrate their ML applications from smaller, locally run resources, such as laptops and workstations, to large-scale HPC systems, such as the SDSC Expanse supercomputer. Participants accelerated their learning process through highly interactive classes with hands-on tutorials on SDSC's Expanse.

HIGH-PERFORMANCE COMPUTING (HPC) STUDENTS PROGRAM

The HPC Students Program focuses on organizing, coordinating and supporting club activities; purchasing/loaning tool and cluster hardware to the club and sponsoring students to travel to the annual Supercomputing Conference (SC). This program also hosts the HPC User Training classes in collaboration with the UC San Diego Supercomputing Club, where participants are taught about the architecture of HPC clusters and learn to run scientific applications on those systems. The program also organizes and awards Co-Curricular Record (CCR) credits to SDSC interns while assisting principal investigators (PIs) to create new CCRs.

STUDENT CLUSTER COMPETITION

Since 2007, the Student Cluster Competition (SCC) has been a featured event at the annual Supercomputing Conference (SC). SCC teams compete against teams from around the world in the non-stop, 48-hour challenge to complete a real-world scientific workload, while keeping the cluster up and running. Acceptance to the competition is stiff and requires intense preparation and skill development. The competing teams consist of a mentor and six students who design and build a small cluster with hardware and software vendor partners, learning designated scientific applications and applying optimization techniques for their chosen architectures. At SC21, SDSC's team finished fourth overall. At SC22, the team won the HPL Benchmark Contest – the first U.S. team to win since 2010 – and placed third overall among 13 teams.

School of Computing, Information and Data Sciences at UC San Diego Marks a New Era

A new School of Computing, Information and Data Sciences (SCIDS) is under consideration at UC San Diego. Envisioned as providing leadership in research, learning and technological developments in the areas of data, information and computing sciences, the new school would align with the university's founding paradigm—to serve as a hub of interdisciplinary inquiry and innovation.

A natural outgrowth of this model, SCIDS would be founded around SDSC and the Halicioğlu Data Science Institute (HDSI). These primary units would garner support from interactions and affiliations with existing university divisions and academic departments, such as computer science and engineering (CSE), electrical and computer engineering (ECE), cognitive science and mathematics.

SDSC would serve as the operational and translational science core, building on its history as one of the original four national supercomputer centers established by the National Science Foundation nearly four decades ago. Since then, SDSC has been a leader in the development of high-performance computing, big data and, more recently, cloud computing and computing continuum. HDSI, created in anticipation of the growth of data sciences and with generous philanthropic support, would serve as the academic core of the new school with its established undergraduate program and approved graduate degree programs.

Beyond HDSI and SDSC, SCIDS would have strong academic interactions involving all UC San Diego departments, schools and divisions supporting the goal of transforming data into knowledge through development of data and information science, advancing innovative computing paradigms and developing entirely new contextual learning algorithms and methodologies that can transform society. The educational programs would be designed to train an entirely new generation of qualified professionals.

To be competitive on the national landscape with recently created schools of similar scope (e.g., the Berkeley School of Computing, Data and Society and MIT's new College of Computing), there would be opportunities for academic units to create formal connections with SCIDS. For example, a formal connection would be established between SCIDS and UC San Diego's Computer Science and Engineering Department as well as the university's Electrical and Computer Engineering Department.

Establishment of the new school is motivated by powerful intellectual and educational goals. The school would provide an approach to synergize the stand-alone academic and research units of HDSI and SDSC in an auspicious manner in the highly competitive world of computing, information and data sciences. As stand-alone units, both HDSI and SDSC are currently overseen by the Senior Associate Vice Chancellor serving in a “Dean-designee” role. The new school would benefit from the appointment of dedicated academic leadership in the form of a new Dean reporting to the Executive Vice Chancellor. This dedicated oversight would position the school to compete successfully in this emerging area.

Beyond administrative streamlining, the proposed SCIDS would open multiple possibilities for new academic programs and research initiatives that the faculty and researchers in the school would be able to draw closely together. This would build on recent successes such as the $20 million AI Institute TILOS, which resulted from interactions such as those the school would promote. It is anticipated that training programs for working professionals, as well as executive training programs in the areas of big data and artificial intelligence, would also be offered through SCIDS.

Currently, the full proposal for the school is pending review by the University of California Office of the President.

SDSC's Evolution

The San Diego Supercomputer Center is one of the nation's premier centers for high-performance and data-intensive computing, and the only center of its kind in the University of California system. The scope in computing and expertise—in scale, nationally and across domains—at SDSC, backed up by expansive computing infrastructure, ongoing grants and contracts, and funded partnerships with industry will immediately catalyze the collaborative research and experiential learning opportunities in SCIDS.

SDSC was established as one of the nation's first supercomputer centers under a cooperative agreement by the National Science Foundation (NSF) in collaboration with UC San Diego and General Atomics (GA) Technologies, opening its doors in 1985. Since then, it has grown and stewarded a national reputation as a pioneer and leader in high- performance and data-intensive computing and cyberinfrastructure. Located on the campus of UC San Diego, SDSC provides resources, services and expertise to UC San Diego, the UC System, the state of California, the national research community and the private sector. SDSC supports a wide range of multi-disciplinary programs that engage tens of thousands of individual researchers and users, spanning a wide variety of domains from astrophysics, biology and earth sciences to bioinformatics and health information technology.

Select dates in the evolution of SDSC are noted below (an extensive, interactive timeline of SDSC's history is available via the link at the bottom of this page).

1985 Founding of SDSC, following award of unsolicited proposal by the founding director, Sid Karin, SDSC opened its doors under a cooperative agreement with General Atomics and UC San Diego. That same year, a Cray X-MP entered production operations as SDSC's inaugural supercomputer.

1997 A partnership led by UC San Diego is one of two winners selected in NSF's Partnerships for Advanced Computational Infrastructure (PACI) competition. As a result, UC San Diego assumes oversight for SDSC, taking over operational responsibility of the center, and transferring all staff from GA employees to UC San Diego employees. At this time the state of California also formalized the broad role of SDSC through line-item funding in the state budget. Over the years, this has evolved from direct funding from the state to funding from the University of California Office of the President (UCOP) via UC San Diego. Today, UCOP funding makes up roughly half of the core budget of SDSC.

2005 NSF awards funding to SDSC as part of the Extensible Terascale Facility (ETF), also called TeraGrid, TeraGrid which at the time, is the world's largest, most comprehensive distributed cyberinfrastructure for open scientific research.

2011 NSF awards funding to SDSC as part of the Extreme Science and Engineering Discovery Environment (XSEDE), the successor to the TeraGrid project. In 2016, NSF extended XSEDE (XSEDE 2.0) another five years, where it remains in operation. Proposals for the XSEDE follow-on are currently under review and we expect SDSC will be part of one or more awards under that program.

2013 UC San Diego and SDSC establish the Triton Shared Computing Cluster (TSCC), a campus computing facility operated via a condominium business model, i.e., researchers buy hardware from a menu of choices offered by SDSC, and SDSC operates the system on behalf of the researchers. UC San Diego provides support for the operating expenses with the understanding that this is more cost effective than researchers deploying hardware in their own buildings. SDSC also offers part of its data center as a UC San Diego-supported co-location facility for hardware owned and operated by UC San Diego researchers, again reducing the overall cost of ownership to the university in terms of space and utilities, while providing better value to the researchers.

2016 – Present In a series of back-to-back awards, SDSC received funding for high-performance computing systems, Gordon, Comet, Expanse, Voyager and the National Research Platform, ensuring SDSC's leadership in supercomputing for the next decade.

Today, SDSC has dozens of principal investigators who obtain extramural funds with expenditures in the millions each year, supporting multiple researchers and staff. Additionally, SDSC has significant and growing education and training programs, as well as several affiliated researchers who teach undergraduate- and graduate-level classes at UC San Diego. SDSC consistently ranks among the top organizational research units by grant funding on campus at UC San Diego.

Timeline of SDSC's History
https://timeline.sdsc.edu/

Turning Words into Action for National Strategy on Integrated, Accessible Data

Scientific research is trending rapidly toward more open, accessible and supportive rapid-response discoveries. At the same time, scientists are collaborating across the U.S. to address complex challenges, such as COVID-19 and supply chain issues.

According to a group of researchers from several universities and institutes across the U.S. and in the Netherlands—including SDSC's Research Data Services (RDS) Director Christine Kirkpatrick and former Associate Director of Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank Philip Bourne—there are robust responses around the world to the need for a unified open research commons (ORC). This is an interoperable collection of data and compute resources within both the public and private sectors that is user-friendly and broadly accessible. And while other nations are gearing up for future competitiveness in this way, the U.S. is lagging behind.

The problem, according to the researchers from places such as MIT, John Hopkins and Argonne National Laboratory, is that the U.S. needs a more committed effort toward making research computing and data infrastructure accessible and connected. Meanwhile, the lag compromises competitiveness and leadership, limiting beneficial U.S. contributions to global science.

"The U.S. has critical mass in experts, forward-thinking program officers and no end to the societal challenges and science use cases that call for a unified research commons, yet it calls for organization at a level higher than these initiatives are usually funded. Immediate and sustained leadership and support in the U.S. are needed to chart the course, starting with policymakers and research funders," said Kirkpatrick, who also leads SDSC's FAIR (findable, accessible, interoperable and reusable) efforts via the U.S. GO FAIR Office located at SDSC.

In an article published last summer in *Science*, the researchers affirm the value of broad cooperation around technology and data. For example, they point to shared governance and infrastructure, as well as standard agreements, that permit a shared system such as the North American electrical grid to direct electricity to where it is needed. They also cite the CIRRUS banking network, which can deliver funds from an individual's bank account to most places around the world. The researchers note that similar coordination in the research enterprise could pay enormous dividends.

"We now have vast amounts of publicly available research data, but to fully leverage the potential power of these data beyond individual and often heroic efforts, these data need to be identified, made interoperable and aligned so that they can be broadly used by the scientific community," said Bourne, first author of the paper, currently with the University of Virginia's School of Data Science, who previously was a professor of pharmacology, and bioinformatics and systems biology at UC San Diego.

According to the researchers, data on disparate topics—such as a county's homelessness rates, average income, neighborhood food and health resources, air pollution, flood risk, predicted water resources and predicted average temperature—often are spread across a range of locations on the web, infrastructures and management regimes.

"If these data were integrated—brought together based on common data elements in each dataset—we could use these data for powerful analyses, like identifying locations with high homeless populations that are also likely to be hit hardest by floods, droughts or heat waves, or places with poor cardiac health that also have high or increasing particulate matter pollution, which could lead to more heart attacks," said Bourne.

Support by policymakers and funders who are driving the development of research infrastructure can facilitate such work, similar to the urgent cooperation seen among scientists during dire times of need, such as the COVID pandemic, the threat of war and the disruption to the global economy.

The researchers hold that the U.S. has a vibrant research ecosystem with no lack of computation and data resources. But, the struggle lies in the cultural and institutional obstacles that require policy leadership and a sustained commitment to overcome.

The approach needed per the researchers is a coherent national strategy that includes: mutually beneficial U.S. industry partnerships, formal executive representation in international ORC-focused initiatives, AI-ready data and long-term data preservation for reproducibility, professional data stewardship and ultimately federal commitment to charting the future and establishing a national ORC. According to the researchers, incentive to create a unified system is paramount.

"Scientists are not yet presented with the adequate incentives. Mandates from funders—such as data-sharing policies—help, but there are not enough definitions of requirements and rewards for complying or, indeed, a unification of what is expected of researchers regardless of the source of their research funding," explained Kirkpatrick, who also serves as secretary general for the International Science Council's (ISC) Committee on Data (CODATA).

Kirkpatrick pointed to some of the efforts SDSC has made toward supporting accessibility and connectedness:

National Science Data Fabric pilot project
SDSC participates in the first infrastructure capable of bridging the gap between massive scientific data sources, the Internet2 network connectivity and an extensive range of HPC facilities and commercial cloud resources around the nation;

The Open Storage Network
A national, distributed storage resource for sharing data at scale;

The National Research Platform
An NSF-funded innovative, all-in-one system that combines computing resources, research and education networks, edge computing devices and other instruments to expedite science and enable transformative discoveries;

CloudBank
An NSF-funded service to help researchers access and use public cloud computing resources;

Open Knowledge Network (OKN)
The NSF-funded, Convergence Accelerator-affiliated program calls for multidisciplinary and multi-sector teams to work together to build a cooperative and shared OKN infrastructure to drive innovation across science, engineering and humanities;

Open Science Chain
A program for providing a secure method to efficiently share and verify data and metadata while maintaining privacy restrictions necessary for the reuse of the scientific data;

Open Science Grid
A consortium that builds and operates a set of pools of shared computing and data capacity for distributed high-throughput computing; and

AI Institute for Intelligent Cyberinfrastructure with Computational Learning in the Environment (ICICLE)
Participation in an institute focused on user-friendly, next-generation intelligent cyberinfrastructure for user-friendly AI applications.

About the collaborative article in *Science*, Bourne noted, "It was wonderful to engage with Christine on this important policy forum and to reengage with SDSC where I spent many happy years. Collectively, we have made an important statement for the future of research computing, and I look forward to helping turn words into action."

SDSC RESEARCHERS

Ilkay Altintas, Ph.D
Chief Data Science Officer
Director, Workflows for Data Science (WorDS) Center of Excellence
Lecturer, Computer Science and Engineering, UCSD

Chaitan Baru, Ph.D.
Senior Science Advisor, Office of Integrative Activities, Office of the Director, NSF
Distinguished Scientist, SDSC
Director, Center for Large-scale Data Systems research (CLDS)
Associate Director, Data Science and Engineering
Associate Director, Data Initiatives

James Bordner, Ph.D.
Senior Computational Scientist

Hans-Werner Braun
Research Scientist
Adjunct Professor, College of Sciences, SDSU
Director/PI, High Performance Wireless Research and Education Network (HPWREN)
Internet Hall of Fame Inductee

Sandeep Chandra, M.S.
Executive Director, Sherlock Cloud
Director, Sherlock Cloud Solutions and Services Division

Dong Ju Choi, Ph.D.
Senior Computational Scientist
Assistant Clinical Professor, Department of Radiation Medicine and Applied Sciences, UC San Diego

Amit Chourasia, M.S.
Senior Visualization Scientist
PI, Stream Encode Explore and Disseminate My Experiments (SEEDME)

Kimberly Claffy, Ph.D.
Director/PI, Center for Applied Internet Data Analysis (CAIDA)
Research Scientist
Adjunct Professor, Computer Science and Engineering, UCSD
Internet Hall of Fame Inductee

Daniel Crawl, Ph.D.
Associate Director, Workflows for Data Science

Yifeng Cui, Ph.D.
Director, High-Performance GeoComputing Laboratory
Director, Intel Parallel Computing Center
PI, Southern California Earthquake Center
Adjunct Professor, SDSU

Diego Davila, M.S.
Computer Scientist

Jose M. Duarte, Ph.D.
Assistant Project Scientist, RCSB Protein Data Bank

Melissa Floca, MBA
Director, Strategic Partnerships, CICORE Division

Anthony Gamst, Ph.D.
Director, Computational and Applied Statistics Laboratory

Andreas Goetz, Ph.D.
Director, Computational Chemistry Laboratory
Co-PI, Intel Parallel Computing Center
Senior Investigator, Center for Aerosol Impacts on Chemistry of the Environment (CAICE), UCSD

Madhusudan Gujral, Ph.D.
Bioinformatics Programmer Analyst

Amarnath Gupta, Ph.D.
Director, Advanced Query Processing Lab of SDSC
Co-PI, Neuroscience Information Framework (NIF) project, Calit2

Bradley Huffaker, M.S.
Senior Research Programmer, CAIDA
Specialist, Computer Networks

Thomas Hutton
Chief Network Architect

Martin Kandes, Ph.D.
Research Specialist, Computational and Data Science

Christine Kirkpatrick, M.A.S.
Division Director, Research Data Services
Head, GO FAIR US
Secretary General, CODATA
PI, EarthCube Office
PI, West Big Data Innovation Hub
Ex Officio Member, U.S. National Committee for CODATA for the National Academics of Sciences, Engineering, and Medicine
Co-Chair, FAIR Digital Object Forum

Valentina Kouznetsov, Ph.D.
Associate Project Scientist
Research Professor

Amit Majumdar, Ph.D.
Division Director, Data Enabled Scientific Computing
PI, Voyager
Associate Professor, Department of Radiation Medicine and Applied Sciences, UC San Diego

Alexander Marder, Ph.D.
Research Scientist, CAIDA

Mark Miller, Ph.D.
PI, Biology
PI, CIPRES gateway
PI, Research, Education and Development Group

Dmitry Mishin, Ph.D.
Applications Developer

Ka Pui Mok, Ph.D.
Research Scientist, CAIDA

Viswanath Nandigam, M.S.
Director (interim), Advanced Cyberinfrastructure Development Lab
PI, OpenTopography
Co-I OpenAltimetry

Mai H. Nguyen, Ph.D.
Lead. Data Analytics

Michael Norman, Ph.D.
Distinguished Professor, Physics, UC San Diego
Director, Laboratory for Computational Astrophysics, UC San Diego

Dmitri Pekurovsky, Ph.D.
Senior Computational Scientist, Scientific Computing Applications Group

Wayne Pfeiffer, Ph.D.
Distinguished Scientist

Zaira Razu, M.A.
Director, Convergence Research (CORE) Institute

Peter Rose, Ph.D.
Director, Structural Bioinformatics Laboratory
Lead, Bioinformatics and Biomedical Applications, Data Science Hub

Joan Segura, Ph.D.
Scientific Software Developer, RCSB Protein Data Bank

Igor Sfiligoi, M.S.
Senior Research Scientist, Distributed High-Throughput Computing
Lead Scientific Software Developer and Researcher

James Short, Ph.D.
Lead Scientist
Co-Director, Center for Large-scale Data Systems Research (CLDS)
Director, BlockLAB

Robert Sinkovits, Ph.D.
Director, Scientific Computing Applications
Director, Education and Training

Subhashini Sivagnanam, M.S
Lead, CyberInfrastructure Solutions and Services
Lead, Triton Shared Computing Cluster
PI, Open Science Chain
Co-PI, Neuroscience Gateway

Shava Smallen, M.S.
Manager, Cloud and Cluster Development
Lead Software Architect and Co-PI, CloudBank
Steering Committee Co-Chair, Pacific Rim Application and Grid Middleware Assembly (PRAGMA)

Shawn Strande, M.S.
Deputy Director

Mahidhar Tatineni, Ph.D.
Lead, User Support
Research Programmer Analyst

Mary Thomas, Ph.D.
Computational Data Scientist
Lead, HPC Training
Co-PI, CC Compute: Triton Stratus*

Igor Tsigelny, Ph.D.
Research Scientist
Research Scientist, Department of Neurosciences, UC San Diego

David Valentine, Ph.D.
Research Programmer, Spatial Information Systems Laboratory

Tanya Wolfson, M.A.
Senior Staff Member, Computational and Applied Statistics Laboratory

Frank Würthwein, Ph.D.
Director
Lead, Distributed High-Throughput Computing
Professor, UC San Diego Department of Physics
Professor, Halıcıoğlu Data Science Institute

Kenneth Yoshimoto, Ph.D.
Researcher, Computational and Data Science

Choonhan Youn, Ph.D.
Scientific Researcher

Ilya Zaslavsky, Ph.D.
Director, Spatial Information Systems Laboratory

Michael Zentner, Ph.D.
Director, Sustainable Scientific Software Division
Director, Science Gateways Center of Excellence (SGX3)
Director, Science Gateways Community Institute (SGCI)

Andrea Zonca, Ph.D.
Specialist, HPC Applications

SDSC@UC San Diego

San Diego Supercomputer Center
University of California San Diego
9500 Gilman Drive MC 0505
La Jolla, CA 92093-0505

www.sdsc.edu
email/info@sdsc.edu
twitter/SDSC_UCSD
instagram.com/SDSC_UCSD
facebook/SanDiegoSupercomputerCenter
linkedin.com/company/san-diego-supercomputer-center
youtube.com/SanDiegoSupercomputerCenter