SAN DIEGO
SUPERCOMPUTER CENTER
UNIVERSITY OF CALIFORNIA SAN DIEGO

Ai
INNOVATION
INSIGHT
IMPACT

ANNUAL REPORT
2023 – 2024

FOR A **BETTER WORLD.**

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, Communications
cdillon@ucsd.edu

INNOVATION, INSIGHT,
IMPACT FOR A BETTER WORLD
SDSC Annual Report
2023 – 2024
(PDF version available online at www.sdsc.edu/pub/)

EDITOR:
Cynthia Dillon

CO-EDITOR:
Kimberly Mann Bruch

CONTRIBUTORS:
Ashley Atkins, Kim Bruch, Cynthia Dillon, Jake Drake, Fritz Leader, Amit Majumdar, Susan Rathbun, Ben Tolo, Frank Würthwein

CREATIVE DIRECTOR:
Ben Tolo

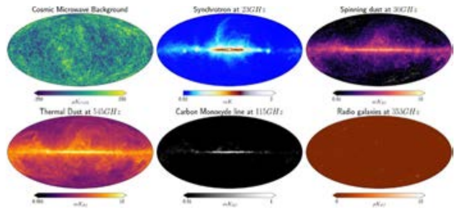
PHOTOGRAPHER:
Jake Drake

All financial information is for the fiscal year that ended June 30, 2023. 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.

About SDSC

Innovation

Insight



Impact

Education & Training

DIRECTOR'S MESSAGE

DIRECTOR'S MESSAGE

INNOVATION INSIGHT IMPACT FOR A BETTER WORLD.

Dear Friends,

During the past year, the San Diego Supercomputer Center has thoughtfully considered how we want to demonstrate leadership in this emergent age of AI. With decades-long expertise in data and cyberinfrastructure, repute in high-performance computing, as well as our growing data center, we are well-positioned to enable cutting-edge advancements that address the most pressing scientific and societal challenges.

SDSC's commitment to AI leadership has been evident for some time. In 2021, for example, we released news about our participation in the AI Institute for Intelligent Cyberinfrastructure with Computational Learning in the Environment—known as ICICLE. This $20 million U.S. National Science Foundation (NSF) initiative is aimed at next-generation intelligent cyberinfrastructure that makes using AI as easy as plugging your cell phone charger into an electrical outlet. ICICLE brings together a multidisciplinary team of scientists and engineers, led by the Ohio State University in partnership with SDSC, UC San Diego, UC Davis and several other universities, institutes and industry partners.

Across its divisions, SDSC offers many tells of its AI leadership strength.

Examples include projects such as the Convergence Research (CORE) Institute and WIFIRE Lab, led out of the Cyberinfrastructure and Convergence Research and Education (CICORE) Division. These teams are comprised of the division's experts in data science, computing, workflow management, GIS, knowledge networks and AI, as well as thematic areas such as earthquakes, wildfires and genomics. Additionally, WorDS Center of Excellence builds upon years of experience developing workflows for computational science, data science and engineering at the intersection of distributed computing, big data analysis and reproducible science.

While these examples serve to outline SDSC's AI leadership, there is color to add. For example, last fall the U.S. Department of Energy announced its selection of a multi-institutional team of data scientists from General Atomics (GA), SDSC, UC San Diego, Hewlett Packard Enterprise and Sapientai to develop a Fusion Data Platform (FDP) for advancing high-priority fusion research, with an award of $7.4 million over three years. Led by GA, the FDP initially will be deployed at SDSC. Once completed, the FDP will be made available to the scientific community to provide access to high-quality fusion data for the efficient creation of reproducible AI and machine learning models to support the design and operation of a broad range of fusion pilot plants (FPP) designs and plasma configurations within a decadal timescale.

Additionally, in terms of data infrastructure, we recently hosted a powerful NDC-C workshop. NDC-C is an NSF program for a "National Discovery Cloud for Climate." It includes more than 30 awards. A majority of the project PIs attended the kick-off event at SDSC. A resourceful outcome of this event was a matrix capturing these connections and easing interfaces with the Open Science Data Federation and the National Data Platform.

Work has continued through the Research Data Services (RDS) Division on the NSF-funded FAIR in Machine Learning, AI Readiness and Reproducibility Research (FARR) Coordination Network. This project brings communities together to foster collaboration, stimulate research and create research roadmaps.

I'll conclude with examples of our AI-related impact for the National Artificial Intelligence Research Resource (NAIRR) Pilot. A team from our Sustainable Scientific Software Division's Cyberinfrastructure Center of Excellence, known as SGX3, led the development of the NAIRR Pilot portal using the Hubzero® framework. It served as a key tool for users during the initial NAIRR allocations call in January 2024.

Additionally, former SDSC director Michael Norman was an original member of the NAIRR Task Force. The NAIRR pilot includes numerous private sector resources; Microsoft will contribute $20 million in cloud credits, which will be distributed through CloudBank, which is led by Norman, at SDSC. These credits are applicable to the spring 2024 NAIRR allocations call. We will also serve as a NAIRR classroom resource through the Prototype National Research Platform (PNRP), and support NAIRR allocations on both Expanse and Voyager. Expanse will be expanded to include 136 NVIDIA H100 GPUs to support the NAIRR Pilot.

The AI climate is electric, and while the future will show us what exactly emerges from it, I know that SDSC will be a constant source of leadership in this space.

Please enjoy catching up with the news and information we generated over the past year as you read through this annual report. Until next year…

Best wishes,
Frank Würthwein,
SDSC Director

SDSC Mission, Vision, Goals

In August 2023, SDSC's Executive Team (ET), comprised of division directors and members of the Director's Office Group, gathered for a two-day retreat. Informed by staff input from efforts in Spring 2023 at the division level to determine team goals and priorities, the ET discussed the future direction of the center. Under the facilitation of Esther Coit from Patera Design, team members worked in small groups and reconvened for full group discussions throughout the retreat. After the final activity, whereby the ET worked in two, larger separate groups, the SDSC leaders met as a whole and discovered that the goals identified by each of the two groups were virtually identical. Following are SDSC's new mission and vision statements, as well as its five strategic goals:

MISSION

Innovating to impact science, technology, education and society

VISION

SDSC will be a global leader in delivering integrated data and computing solutions that enable translational research, diverse partnerships and workforce development.

GOALS

1. **Leadership in Integrative AI, Data and Computing Solutions:** Develop and deliver transformational applications through innovative research.
2. **Workforce Growth:** Expand and diversify our team of experts.
3. **Financial Sustainability:** Practice sustainable and scalable financial management.
4. **Culture Revitalization:** Foster an inclusive, supportive workplace where employees excel and collaborate to meet our goals, and where stakeholders engage with us.
5. **Education:** Implement an equitable and strategic approach to educating and training our stakeholders.

Office of the Director's Group

Frank Würthwein, Ph.D.
Director

Frank Würthwein is the director of SDSC. Würthwein leads Distributed High-Throughput Computing at SDSC, and he is a faculty member in the UC San Diego Department of Physics, as well as a founding faculty member of the Halicioğlu Data Science Institute on campus. His research focuses on experimental particle physics, and in particular the Compact Muon Solenoid experiment at the Large Hadron Collider. He continues to serve, as he has for many years, as executive director of the Open Science Grid, the premiere national cyberinfrastructure for distributed high-throughput computing.

Fritz Leader
Chief Administrative Officer

Fritz Leader is the chief administrative officer of SDSC and a member of the executive team. He leads SDSC's Business Office in its support of finances, human resources and facility space. Leader is a 20-year employee of UC San Diego, and he has held his role at SDSC since 2017.

Rick Wagner, Ph.D.
Chief Technology Officer

Rick Wagner is the chief technology officer of SDSC, where he formerly served as HPC systems engineer and then HPC systems manager between 2010 and 2016. He joined the University of Chicago as a member of the Globus management team and supervised the Professional Services group, working closely with the research community to design and implement large-scale data initiatives. In 2020, Wagner returned to UC San Diego as principal systems integration engineer in the Research IT Services group, helping campus faculty and staff adopt new technologies and services.

Ashley Atkins, M.S.
Chief of Staff

Ashley Atkins is the chief of staff of SDSC. She most recently served as the executive director of the NSF-funded West Big Data Innovation Hub at UC Berkeley. Through her research, she has focused as a PI and Co-PI on water resource modeling and societally beneficial applications of data. She is a Fulbright alumna.

SDSC Executive Team

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

Ilkay Altintas
Chief Data Science Officer
Division Director, Cyberinfrastructure and Convergence Research and Education Division

Ashley Atkins
Chief of Staff

Sandeep Chandra
Division Director, Sherlock

Cynthia Dillon
Division Director, Communications

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
Deputy Director

Rick Wagner
Chief Technology Officer
Director, Industry Partnerships

Michael Zentner
Division Director, Sustainable Scientific Software

Divisions of SDSC

SDSC is organized into seven divisions. Each operates independently under the guidance of a director, but as a vital component of the collective whole, working together to meet the mission, vision and goals of SDSC.

BUSINESS OFFICE

The Business Office is the administrative hub of SDSC. In addition to administrative support, the staff provides the entire center with assistance related to finance, human resources, facilities and data systems.

Director: Fritz Leader

CYBERINFRASTRUCTURE AND CONVERGENCE RESEARCH AND EDUCATION

The Cyberinfrastructure and Convergence Research and Education (CICORE) Division combines broad expertise in cyberinfrastructure with deep, domain-specific expertise in AI-enabled science to build use-inspired solutions to grand societal challenges at scale with partners in research, practical and industry communities.

Director: Ilkay Altintas

COMMUNICATIONS

Originally referred to as External Relations, the renamed Communications Division serves as an internal public relations unit for SDSC. It promotes and applies strategic communications expertise and best practices in support of SDSC's brand. The team works to be the driving force behind SDSC's recognition as a leader of innovation and societal impact, amplifying SDSC's voices and stories to inspire, connect and empower its communities.

Director: Cynthia Dillon

DATA-ENABLED SCIENTIFIC COMPUTING

The Data-Enabled Scientific Computing (DESC) Division is organized into multiple groups that have specific expertise to lead HPC and computational sciences innovation. The division serves and collaborates with thousands of SDSC's national, UC-wide and UC San Diego researchers, as well as with industry partners to provide full support and training to SDSC's user community.

Director: Amit Majumdar

RESEARCH DATA SERVICES

RDS, as it is known at SDSC, provides services that enable researchers to attain their research and computing goals—from power and network systems to systems integration. Division staff serve the varying needs of researchers—cloud computing, storage for active workloads or archival use cases, as well as backup storage for disaster recovery—always collaborating and anticipating ways to meet evolving research needs.

Director: Christine Kirkpatrick

SHERLOCK

This division works to solve the mysteries that cyberinfrastructure and cloud computing can present to the user communities it serves. Sherlock's core expertise lies at the intersection of cloud computing, cybersecurity and regulatory compliance. In support of these key areas the division is subdivided into Cloud Architecture and DevOps, Cloud Infrastructure, Data Architecture and Platforms, Cybersecurity and Compliance, and Outreach and User Support groups.

Director: Sandeep Chandra

SUSTAINABLE SCIENTIFIC SOFTWARE

The SD3 Division addresses the challenge for scientists of creating and managing computational resources by providing services in several areas: cyberinfrastructure, software and project sustainability planning, professionalized software development and operations, and next-generation tools for biology. The division also has a strong bioinformatics team, including the CIPRES science gateway framework team, the Hubzero platform and SGX3.

Director: Michael Zentner

SDSC BY THE NUMBERS

FISCAL YEAR 2022/23

ORGANIZATION	
$64M	ANNUAL REVENUE
$36M	GRANT FUNDING
$8M	INDUSTRY REVENUE
293	EMPLOYEES & VOLUNTEERS
57	SPONSORED RESEARCH AWARDS

TRAINING PROGRAMS	
40,570	STUDENTS ENROLLED WORLDWIDE IN SDSC-LED ONLINE COURSES
2,908	TRAINING & EVENT PARTICIPANTS
272	STUDENTS ENGAGED IN SDSC PROGRAMS
200	HIGH SCHOOL STUDENTS MENTORED
60	SDSC HOSTED PROGRAMS

SUPPORT FOR UC	
2,339	UC SAN DIEGO USERS: EXPANSE, VOYAGER, NRP & COMET
741	UC SAN DIEGO USERS: TRITON SHARED COMPUTING CLUSTER
651	UC SAN DIEGO ACTIVE ALLOCATIONS: EXPANSE, VOYAGER, NRP & COMET
50+	PETABYTES OF STORED DATA FOR UC SAN DIEGO & UC

HPC SYSTEMS	
244K	x86 CORES ON SDSC HPC SYSTEMS
3,014	ACCELERATORS
605	PUBLICATIONS CITING SDSC/ACCESS RESOURCES
6	COMPUTE CLUSTERS

Mahidhar Tatineni: SDSC's Pi Person of the Year

Mahidhar Tatineni (PhD, UCLA) has been named the San Diego Supercomputer Center (SDSC) 2024 Pi Person of the Year. This award, first bestowed to an SDSC researcher in 2013, recognizes an individual whose impact straddles both science and cyberinfrastructure technology.

As SDSC's User Services group lead and a computational data science research specialist manager, Tatineni has completed many optimization and parallelization projects of scientific codes and benchmarks with the center's supercomputing resources. His impact at SDSC for over nearly two decades has been demonstrable in terms of his involvement with U.S. National Science Foundation (NSF)-, Department of Defense (DoD)- and industry-funded research in high-performance computing (HPC), high-throughput computing (HTC), accelerators, other cyberinfrastructure (CI) areas and domain science simulations. His involvement in such efforts as a principal investigator—PI or co-PI—has impacted and enabled research for tens of thousands of national researchers and students who use SDSC's resources for conducting science and education activities. Tatineni has led the creation of new materials and teaching topics on HPC/HTC/accelerators and CI, and he has presented at training sessions, workshops and summer institutes at SDSC, as well conducting tutorials at national competitive venues such as the International Conference for High Performance Computing, Networking, Storage and Analysis (SC) and Practice and Experience in Advanced Research Computing (PEARC).

During the last five years, Tatineni has guided students as a mentor for the Supercomputing Student Cluster Competition at SC. He has served in national-scale reviews and on committees such as NSF proposal review panels, SC/PEARC and journal reviews, allocation review committees and NSF's COVID consortium panel. Additionally, he contributes to the ICICLE AI Institute's CI architecture, especially as it relates to HPC resources at SDSC.

As a member of the Data-Enabled Scientific Computing (DESC) team, Tatineni has demonstrated the deepest and broadest combined impact in terms of research in CI and enabling CI/HPC for many thousands of researchers and students through SDSC's Education Outreach and Training programs, according to DESC Division Director Amit Majumdar.

"His intellect, expertise and contributions have enabled SDSC to acquire several grants for supercomputers—Gordon, Comet, Expanse, Voyager, Prototype National Research Platform (PNRP) and a new system set to launch soon, as well as for cyberinfrastructure research totaling nearly 100 million dollars over the last 15 years," said Majumdar. "Mahidhar is dedicated to his work and as a result has made tremendous impact in CI and research for SDSC."

Tatineni has made many contributions to fundamental HPC research over the years. He has helped design and implement middleware on many of SDSC's supercomputers, and he has optimized, benchmarked and analyzed performance of scientific applications from domain sciences of computational fluid dynamics (CFD), neuroscience, bioinformatics, molecular dynamics and more. For example, his implementation and improvement of scalability of a parallel petascale kinetic magnetosphere simulation code on the Blue Waters machine at NCSA resulted in more than 260 journal citations. He also co-authored a paper that resulted in the Gordon Bell Special Prize for HPC-Based COVID-19 Research at SC20. His library of accepted grants shows the depth and breadth of his research contributions.

Team members note that Tatineni has used many large scale national academic HPC platforms in American HPC during his career. He started as a grad student in the 1990s as an early user of the Maui High Performance Computing Center IBM SP2 supercomputer, continued with large scale runs on NSF machines such as Kraken (the first academic supercomputer to break the petaflop barrier) at the University of Tennessee, Ranger, Stampede2, and Frontera at TACC, Blue Waters at NCSA and all of the systems at SDSC since 2005. He has also participated in MVAPICH2 software implementation and benchmarking on SDSC's HPC resources and other software infrastructure projects, such as HADOOP, Big Data software stack and AI software stack.

According to Majumdar, SDSC researchers and staff who report to him, as well SDSC leaders and his peers within SDSC and outside of SDSC respect him, making Tatineni a "perfect recipient" of the 2024 PI Person of the Year award.

SDSC was established as one of the nation's first supercomputer centers under a cooperative agreement by the NSF in collaboration with UC San Diego and General Atomics (GA) Technologies. SDSC first opened its doors on Nov. 14, 1985, and today its relationships with the NSF and GA remain strong and vibrant.

For nearly 40 years, SDSC has cultivated its national reputation as a pioneer and leader in high-performance and data-intensive computing and cyberinfrastructure. Unique to the University of California system, SDSC provides its resources, services and expertise to local, regional and national partners in academia and industry.

As an Organized Research Unit at UC San Diego, SDSC's edge in cyberinfrastructure centers precisely around an accessible and integrated network of computer-based resources and expertise—focused on accelerating scientific inquiry and discovery. With Expanse and Voyager, SDSC's newest supercomputing resources, the center supports hundreds of multidisciplinary programs spanning a wide range of science themes—from astrophysics and bioinformatics to earth sciences and health IT.

Inspired by a theme of "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.

SERVICES

SDSC offers a variety of research computing cyberinfrastructure resources, services and expertise. The SDSC Data Center houses a wide range of computational resources that are available to UC San Diego, all UC campuses and partner institutions. SDSC services include:

HIGH-PERFORMANCE COMPUTING – SDSC 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, including 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 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. The team offers a Helpful Tools training in addition to answering specific questions via the SDSC Consulting group;

TRAINING – SDSC supports users of its advanced computing systems, including industry and K-14 users, with education and training programs and events such as conferences, panel discussions, symposia, workshops and two summer institutes.

SDSC Researchers

Ilkay Altintas, Ph.D.
Chief Data Science Officer
Division Director, Cyberinfrastructure Research, Education, and Development (CICORE)
Director, Workflows for Data Science (WorDS) Center of Excellence
Director, WIFIRE Lab
Lecturer, Computer Science and Engineering, UC San Diego

Ashley Atkins, M.S.
Chief of Staff

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, UC San Diego
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.
Computational Data Science and Research Specialist

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

Sandra Gesing, Ph.D.
Senior Researcher
Executive Director, U.S. Research Software Engineer Association

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), UC San Diego

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, 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

Rodman Linn, Ph.D.
Associate Director, Fire Science of the WIFIRE Lab

Timothy Mackey, Ph.D.
SDSC Affiliate
Professor, Global Health Program, UC San Diego
Director, Global Health Policy and Data Institute

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

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, 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

Wayne Pfeiffer, Ph.D.
Distinguished Scientist

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

Paul Rodriguez, Ph.D.
Research Analyst

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)

Claire Stirm, M.S.
Project Manager, Science Gateways Center of Excellence (SGX3)

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

Rick Wagner, Ph.D.
Chief Technology Officer
Director, Industry Partnerships

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

EXPANSE

Expanse supports the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) core software stack, which includes remote login, remote computation, data movement, science workflow support and science gateway support toolkits.

As an ACCESS computing resource, Expanse is accessible to ACCESS users who are given time on the system. To obtain an account, users may submit a proposal through the ACCESS Allocation Request System. Interested parties may contact the ACCESS Help Desk for assistance with an Expanse proposal.

Full instructions can be found by visiting the Expanse User Guide page on the SDSC website.

Expanse User Guide

www.sdsc.edu/support/user_guides/expanse.html

Expanse

Expanse is a dedicated Advanced Cyberinfrastructure Coordination Ecosystem: Services and Support (ACCESS) cluster designed by Dell and SDSC, delivering 5.16 peak petaflops and offering Composable Systems and Cloud Bursting. It is a U.S. National Science Foundation-funded system operated by SDSC and available for use through the ACCESS program.

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 (32 GB SMX2) connected via NVLINK and dual 20-core Intel Xeon 6248 CPUs. Expanse also has four 2 TB large memory nodes.

The supercomputer is organized into 13 SDSC Scalable Compute Units (SSCUs), comprising 728 standard nodes, 54 GPU nodes and four large-memory nodes. Every Expanse node has access to a 12 PB Lustre parallel file system (provided by Aeon Computing) and a 7 PB Ceph Object Store system. Expanse uses the Bright Computing HPC Cluster management system and the SLURM workload manager for job scheduling.

RESOURCE ALLOCATION

- The maximum allocation for a PI on Expanse is 15M core-hours and 75K GPU hours. Limiting the allocation size means that Expanse can support more projects, since the average size of each is smaller.
- Science Gateways users in the Maximize tier can request up to 30M core-hours.

JOB SCHEDULING

- The maximum allowable job size on Expanse is 4,096 cores—a limit that helps shorten wait times since there are fewer nodes in idle state waiting for large number of nodes to become free.
- Expanse supports long-running jobs—run times can be extended to one week. Users requests will be evaluated based on number of jobs and job size.
- Expanse supports shared-node jobs (more than one job on a single node). Many applications are serial or can only scale to a few cores. Allowing shared nodes improves job throughput, provides higher overall system utilization and allows more users to run on Expanse.

Expanse in the News

Following are some headline highlights from articles demonstrating how Expanse advances research via a range of scientific domains and uses.

Read Full Articles

www.sdsc.edu/news_and_events

Using Supercomputers to Help Design Quality Recyclables from Plastic

Disposing of plastic products can pose challenges and present risks to ecosystems and human health, but supercomputer models offer promising solutions to the problem.

Published April 24, 2024

Expanse Simulations Illustrate Combo-Antibiotic Plans for TB Treatments

Researchers use quick and economical computational models to configure appropriate treatments for tuberculosis patients.

Published March 28, 2024

Researchers Use Expanse for Project Supportive of Efforts toward Clean Energy

Calculations show how carbon dioxide from fossil fuel combustion can be converted to higher-carbon chain fuels–supporting the nation's transition to clean energy.

Published January 24, 2024

SDSC's Expanse Used for Advancing Undergraduate Experiences in Classroom and Lab

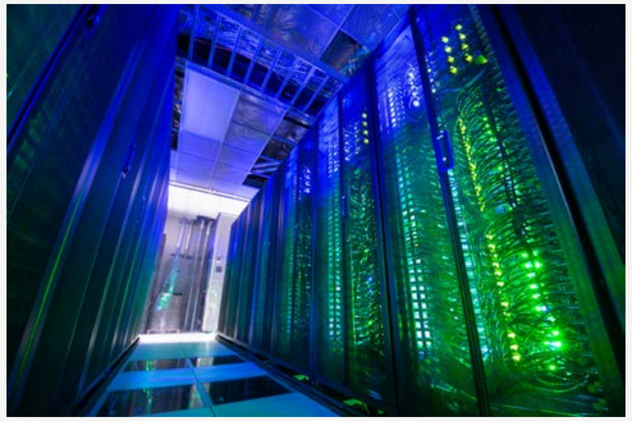
Chemistry faculty and students with limited access to top-of-the-line HPC resources at their local university used Expanse and presented their findings at the International Symposium on Small Particles and Inorganic Clusters.

Published December 11, 2023

Computer Simulations Reveal New Information about Key Protein's Role in Treatment of Heart, Kidney and Other Diseases

The AT1 receptor is a very important drug target. Understanding how physical stimuli affect its function may help the scientific community develop better treatments for heart and kidney diseases.

Published October 5, 2023



Avocado Genome is at the Center of New Study

Researchers turned to a descendant variety of Hass—the world's most popular avocado—to explore the fruit's evolutionary history through its genome.

Published September 5, 2023

Supercomputer Simulations Generate Surge of New Data Related to Stormy Weather Predictions

Using SDSC's Expanse, researchers uncovered new insights into the role of convection and its importance to thunderstorms and weather systems in energy distribution at both local and global scales.

Published July 27, 2023

Groundbreaking Simulations of Water's Phase Diagram Advance Study of Complex Molecular Systems

In a study published in Nature Communications, scientists from UC San Diego made a major stroke in simulating water's phase diagram with remarkable precision using SDSC's Expanse supercomputer.

Published June 13, 2023

This research won the HPCwire Editors' Choice Award for Top HPC-Enabled Scientific Achievement.

CloudBank Resources Raise Data Science Education Programs to New Heights

CloudBank is an integrated service provider that functions to broaden access and impact of cloud computing across the many fields of computer science research and education. It is supported through an award funded by the U.S. National Science Foundation (NSF) to UC San Diego, UC Berkeley and the University of Washington.

Because computational needs often fluctuate throughout university-level courses, one of CloudBank's goals has been to facilitate the use of public cloud resources in the classroom. Simultaneously, UC Berkeley's College of Computing, Data Science and Society developed the Berkeley Data Stack—an auto-scaling, Jupyter hub-based learning platform that scales up when assignments are due and then scales back down in between assignments.

As a CloudBank partner, UC Berkeley wanted to share this platform with instructors at community colleges that may not have the local resources to set up and support a similar system. Berkeley's goal was to reach approximately 500 students among six community colleges through its partnership with SDSC at UC San Diego.

Led by Co-Principal Investigator Shava Smallen, the CloudBank team at SDSC provided administrative support for the project. Smallen said she worked with Berkeley's Kathy Yelick, Eric Van Dusen and Sean Morris as they deployed their educational teaching stack—including a Jupyter hub, a set of labs and notebooks, interactive links for accessing the content and an autograding solution.

"CloudBank makes it very easy for new learners to start computing on their own as all of the course materials are accessed through a browser window with no installation," said Van Dusen, outreach lead for Berkeley's College of Computing, Data Science and Society. "Having a CloudBank Juptyer hub is helping to put community college education on the same foundation as UC Berkeley with the same infrastructure for interactive computing."

Van Dusen said that the team has worked with a non-profit infrastructure organization called 2i2c for implementing the pilot at the following community colleges:

- El Camino Community College, Los Angeles
- Santa Barbara Community College, Santa Barbara
- City College of San Francisco, San Francisco
- Palomar Community College, San Diego
- Skyline Community College, San Mateo
- San Jose Community College, San Jose

"The use of CloudBank allows us to use larger datasets than we would be able to use with our local machines," said Peter Chen, Palomar College instructor. "We can experiment with various data tools that aren't realistic without the use of cloud computing. We are really happy with this partnership and thank the team at Berkeley and SDSC for the opportunity."

The UC Berkeley platform was initially developed for the popular Data 8 course on the campus and quickly expanded to many other courses at UC Berkeley, using local resources to support 10,000 students per month.

HPCwire
2023
READERS' CHOICE AWARDS
Best Use of
HPC in the Cloud
(Use Case)

San Diego Supercomputer Center at UC San Diego, UC Berkeley, and the University of Washington worked together on CloudBank to enable the use of public cloud resources to deploy UC Berkeley's educational teaching stack, including a Jupyter hub, to classrooms at community colleges that may not have the local resources to set up and support a similar system.

Awards Winner for 12 Years
★★★★★★★★★★★★

CloudBank Honored with HPCwire Award

In the 20th Anniversary edition of the HPCwire Readers' Choice Awards, presented at the 2023 International Conference for High Performance Computing, Networking, Storage, and Analysis (SC23), in Denver, CO, in November 2023, SDSC was among the winners. For SDSC's role in the CloudBank project with UC Berkeley, HPCwire recognized the center with a **Readers' Choice Award for Best Use of HPC in the Cloud (Use Case)**.

SDSC also received an **Editors' Choice Award for Top HPC-Enabled Scientific Achievement**. This award recognized multiple ACCESS resources (SDSC's Expanse and the National Center for Supercomputing Applications' Delta), as well as numerous other computational resources, which UC San Diego researchers used to accurately model water phases. This research outcome has tremendous potential to impact computational molecular science and revolutionize scientists' understanding of the molecular world.

Hubzero

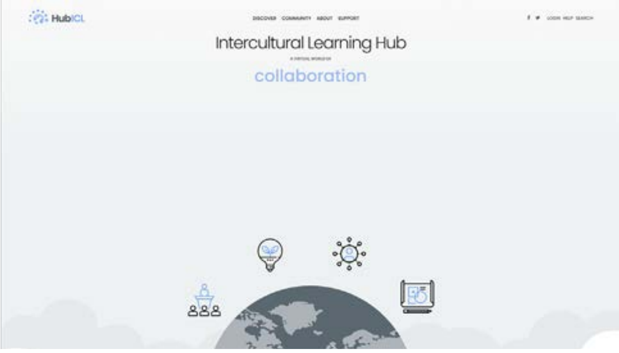
Hubzero is an open source software platform for building powerful websites that host analytical tools, publish data, share resources and collaborate to build communities in a single web-based ecosystem. Initially created by researchers in the NSF-sponsored Network for Computational Nanotechnology to support nanoHUB.org, the Hubzero platform now supports science gateways from a variety of disciplines with a collective of over two million visitors each year.

Hubzero includes a powerful content management system built to support scientific activities. Members on a hub can write blog entries, participate in discussion groups, work together on projects, publish datasets and computational tools with digital object identifiers (DOIs), and make these publications available for others to use—not as dusty downloads, but as live, interactive digital resources. Simulation/modeling tools published on a hub can be accessed with the click of a button, running on cloud computing resources, campus clusters and other national HPC facilities to serve up compelling visualizations.

Hubzero partners enables researchers to:

- Create datasets and interactive simulation tools using RStudio, Jupyter and other web applications;
- Publish research products including, datasets, tools and white papers—through a step-by-step guided system; and
- Provide spaces for research teams and collaborators to discuss data concepts, track progress and share files by using familiar services like Google Drive, GitHub and Dropbox.

Hubzero is a combined effort of the Sustainable Scientific Software Division (S3D) at SDSC, working with the Research Data Services and Sherlock Divisions at SDSC. The core Hubzero team also collaborates with S3D's SDx professional software development and operations group.



Purdue University's Intercultural Hub (HubICL) is one recent example of the many gateways housed by the Hubzero team at SDSC.

Prototype National Research Platform

Having successfully completed its acquisition review last May, PNRP—an innovative system funded by the U.S. National Science Foundation (NSF) and created to advance scientific discoveries—has been operating as a testbed for the past year with two more years of the phase to go. During these testbed years, researchers are exploring the system's design and hardware for use in science and engineering research. Innovative features of the platform include field programmable gate arrays (FPGAs), composable infrastructure and graphics processing units (GPUs). Following the testbed phase, PNRP will become broadly available through a formal allocation process.

"Reaching the milestone was a culmination of a multi-year process from proposal, through acquisition, deployment, early user operations and formal review. It meant the attainment of our goal to provide the research community with an open system created for growth and inclusion; a way for academic institutions to join and participate in a national system to enlarge and enrich the national cyberinfrastructure ecosystem," said SDSC Director Frank Würthwein.

As a distributed system, PNRP features hardware at three primary sites—SDSC, the University of Nebraska-Lincoln (UNL) and the Massachusetts Green High Performance Computing Center (MGHPCC). In addition to the computing hardware at each of the primary sites, the system includes five data caches that are collocated and distributed on the Internet2 network backbone. The data caches provide data replication and movement services that reduce the round trip latencies from anywhere in the U.S. to about 10 milliseconds, or 0.01 seconds.

"The PNRP collaboration represents the future of distributed research computing, where the sources and users of data are part of an integrated fabric. We are excited to support this next phase of the project and look forward to working with the PNRP team over the coming years to realize a vision of enabling research data, anywhere, any time," said James Deaton, vice president of Network Services at Internet2.

Reliability testing of the system has been run to identify any problems that in rare instances might occur, or that become apparent only when running at scale. According to PNRP administrators, the tests showed that PNRP hardware at each of the facility sites performed well.

"One of the most interesting features of the PNRP is the distributed systems management model," said Derek Weitzel, who leads UNL's responsibility for systems administration in the new platform. "PNRP was integrated into existing infrastructure that had been developed over the past several years. The Kubernetes-based approach substantially reduced the time to deploy and integrate hardware. UNL received the cluster on a Monday and had jobs running on Friday that same week, something that would be nearly impossible with a traditional HPC cluster."

John Goodhue is the executive director of MGHPCC, which is operated by a consortium of universities in the northeast, serves thousands of researchers locally and around the world and houses one of the PNRP GPU resources—providing a full complement of data center facility, networking, security and 24/7 operations. "We are pleased to be collaborating on PNRP, which, like MGHPCC, seeks to strengthen the national CI ecosystem through regionally based partnerships," Goodhue said. "PNRP is innovative in technological and organizational dimensions, both of which are essential ingredients to advancing research."

EARLY-USER FEEDBACK

PNRP underwent a 30-day Early User Operations phase, during which the system was put through its paces on real-world applications in preparation for operations. Early-use cases ranged from studies on autonomous agents (e.g., robots, drones and cars) and cerebral organoids to synthesizing textures for 3D shapes and estimating sea surface temperature in cloudy conditions. Early users included researchers from UC campuses, MIT, the International Gravitational Wave Network and additional institutions.

SUPPORT FROM INDUSTRY

Industry partners provide key technical features of the HPC subsystem, which include a mix of FPGA chips, GPUs with memory and storage in a fully integrated extremely low-latency fabric from GigaIO, which provides the composable architecture of the new platform. PNRP's high-performance, low latency cluster integrated by Applied Data Systems (ADS) features composable PCIe fabric technology, along with FPGAs and FP64 GPUs, and two A10-based GPU clusters integrated by Supermicro, one located at UNL and one at MGHPCC.

According to GigaIO, composability provides users with flexibility and the ability to use accelerators such as GPUs and FPGAs in an easy-to-orchestrate, reconfigurable system that saves time and makes optimal use of the resources. "The ability to build formerly impossible computing configurations and seamlessly transform systems to match workloads enables customers like SDSC to do more science for less money. We are proud to have worked closely with SDSC, ADS and Gigabyte to bring this revolutionary system online and make it available to all PNRP researchers," said Alan Benjamin, CEO of GigaIO.

ADS President Craig Swanson said that it was an honor to be selected as the integration vendor partner to build, configure and support the cutting edge composable infrastructure. "It's only our ability to execute and work closely with our partners, that we are able to stand up such bleeding-edge technology to aide in the research community's quest to push the boundaries of science," he said.

PNRP Early Use Cases

IceCube Neutrino Observatory

IceCube is located at the South Pole and consists of 5,160 digital optical modules (DOMs) distributed over one km3 of ice. Determining the direction of incoming neutrinos depends critically on accurately modeling optical properties of the ice. This numerically intensive process needs up to 400 GPU years and a new model must be constructed annually to account for ice flow.

The observatory's computing director, Benedikt Riedel, said, "PNRP's usability was very good and porting efforts were minimal, with only storage needing to be accessed differently and the computation appearing like any other Open Science Grid (OSG) site," adding that performance of the A10 GPUs was excellent.

Credit: IceCube/NSF

Genomics Processing and Analysis

UC San Diego's Tianqi Zhang and Tajana Rosing, one of the PNRP co-principal investigators, developed applications that run on FPGA accelerators for basic genomics processing components, like sequence trimming and alignment, and integrated them with the pipelines for COVID-19 phylogenetic inference, microbial metagenome analysis and cancer variant detection.

"It's pretty easy to migrate the previous programs to the new U55C cluster [PNRP]. The development platform is also similar to the local environment, with only a few board configurations needing administrator intervention. We are currently scaling up and optimizing the accelerators on the multi-FPGA nodes. If successful, it will provide O(10x) speedup and O(100x) power savings compared to CPU," said Zhang.

According to Robert Sinkovits, an expert in scientific applications at SDSC, with the variety and scale of applications and use cases, SDSC "feels confident the [scientific] community will be able to make excellent use of PNRP."

SDSC Contributes to Second Round of NAIRR Pilot Resources

In May 2024, the U.S. National Science Foundation (NSF) and the Department of Energy announced the first 35 projects that will be supported with computational time through the National Artificial Intelligence Research Resource (NAIRR) Pilot, marking a significant milestone in fostering responsible AI research across the nation. The initial call for applicants was issued in January 2024.

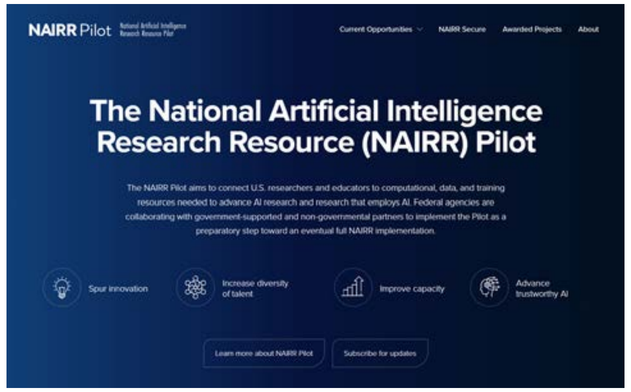
Along with publicizing the news about the first projects, the NSF opened the next opportunity for researchers and educators to apply for access to resources that support AI research. Such resources include advanced computing systems; cloud computing platforms; access to foundation models, software and privacy enhancing technology tools; collaborations to train models and education platforms—areas of strength for SDSC, which is one of the university providers for both research and education resources.

The opportunity also opened up cutting-edge resources contributed by the NAIRR Pilot's nongovernmental partners, including Microsoft, Amazon Web Services, NVIDIA, SambaNova Systems, Cerebras, OpenAI, Anthropic, Groq, EleutherAI, OpenMined, Hugging Face and Vocareum.

“We are excited about the breadth of contributions by SDSC to the second round of NAIRR Pilot resources, including Expanse and Voyager as compute resources, CloudBank to manage the AWS and Azure cloud credits and the Prototype National Research Platform (PNRP) as a NAIRR Classroom resource,” said SDSC Director Frank Würthwein. “We are very much looking forward to an exciting evolution of the NAIRR Pilot program as all of these new diverse sets of resources are being added.”

With this second opportunity, the NSF seeks to connect educators and instructors in universities to computing, data and software resources that will enable them to train their students through hands-on projects and exercises. SDSC is one of presently only two providers for NAIRR Classroom resources. Researchers and educators can apply for access to these resources.

According to NSF Director Sethuraman Panchanathan, the announcement of the first round of NAIRR Pilot projects marks a pivotal moment in the advancement of AI research. “The NAIRR Pilot, fueled by the need to advance responsible AI research and broaden access to cutting-edge resources needed for AI research, symbolizes a firm stride towards democratizing access to vital AI tools across the talented communities in all corners of our country,” said Panchanathan. “While this is only the first step in our NAIRR efforts, we plan to rapidly expand our partnerships and secure the level of investments needed to realize the NAIRR vision and unlock the full potential of AI for the benefit of humanity and society.”



The National Artificial Intelligence Research Resource (NAIRR) Pilot

The NAIRR Pilot aims to connect U.S. researchers and educators to computational, data, and training resources needed to advance AI research and research that employs AI. Federal agencies are collaborating with government-supported and non-governmental partners to implement the Pilot as a preparatory step toward an eventual full NAIRR implementation.

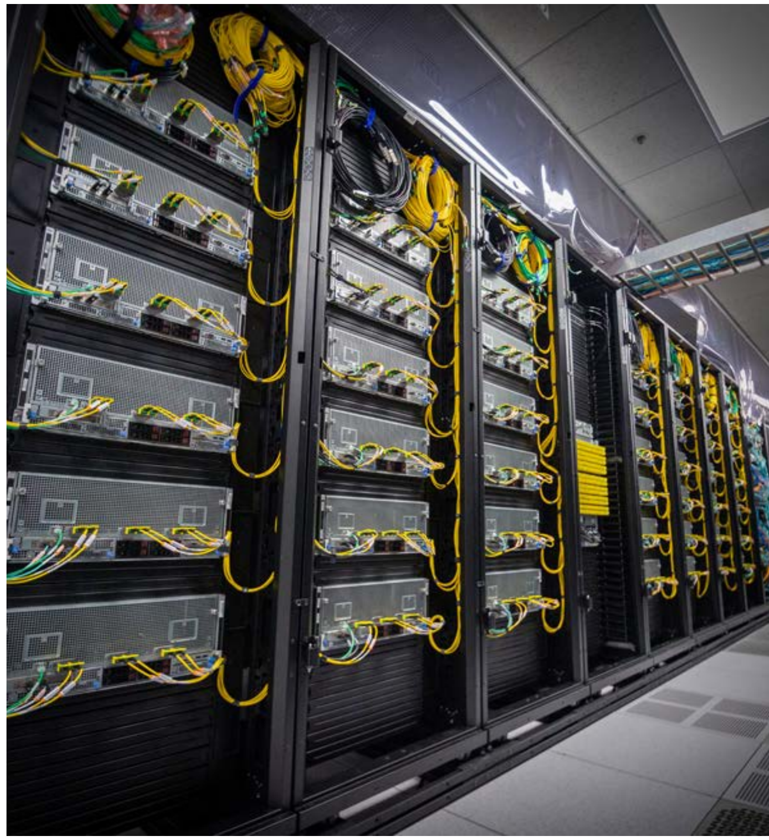
Projects granted computing allocations in the initial round encompass a diverse range of AI-related areas, including investigations into language model safety and security, privacy and federated models, and privacy-preserving synthetic data generation. Other projects also focus on domain-specific research, such as using AI and satellite imagery to map permafrost disturbances, developing a foundation model for aquatic sciences, securing medical imaging data, and using AI for agricultural pest identification.

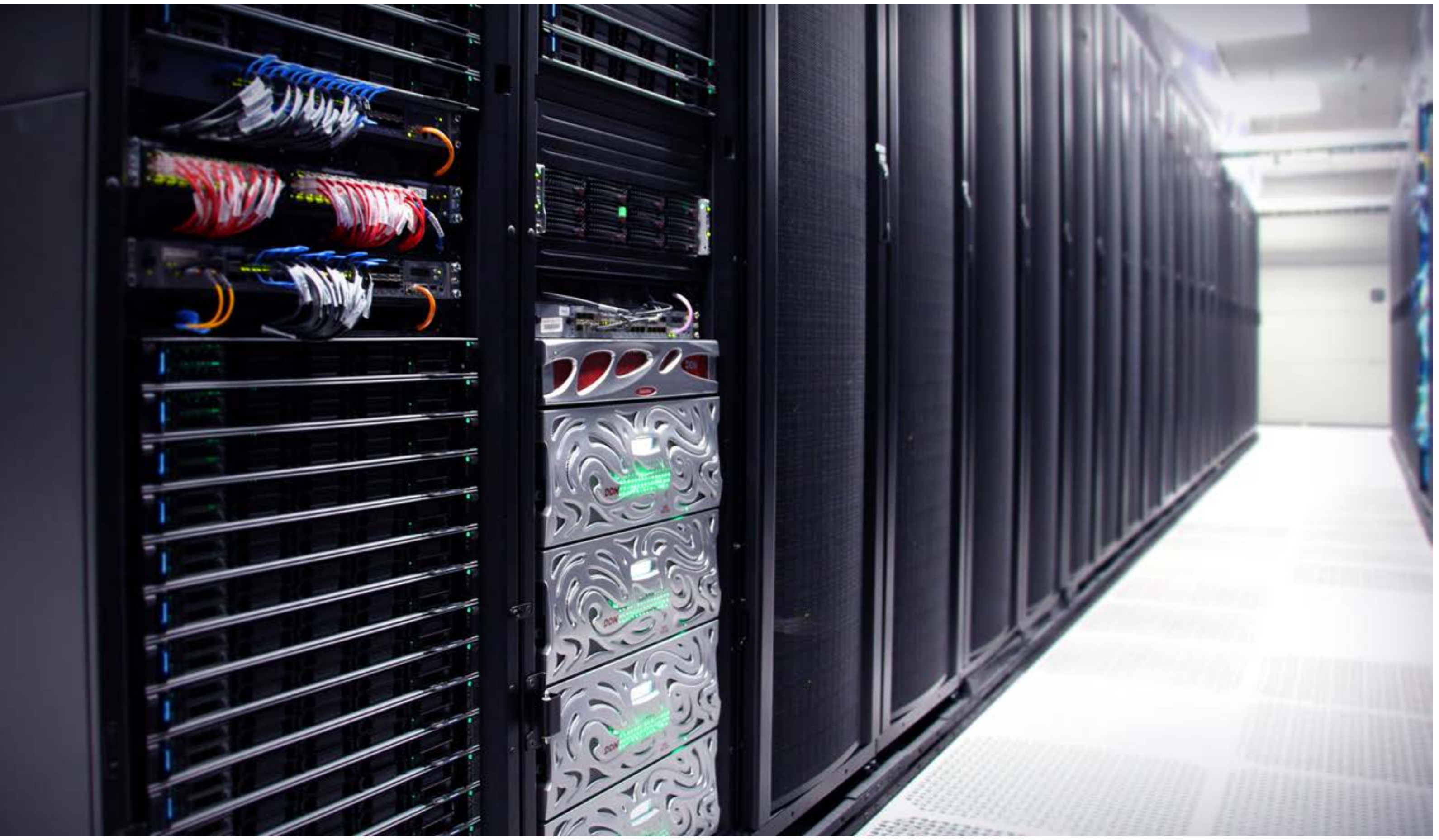
“We look forward to integrating the NAIRR Pilot into the cloud credits distributed and managed by CloudBank,” said Shava Smallen, SDSC's co-principal investigator of CloudBank, a cloud access entity designed to simplify access to emerging computational resources through managed services. “It significantly expands the existing AI focus of what CloudBank provides to the community, and it diversifies the communities CloudBank serves beyond researchers with Computer and Information Science and Engineering NSF awards.”

Voyager

Voyager is an Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) innovative AI system designed specifically for science and engineering research at scale. Voyager is focused on supporting research in science and engineering that is increasingly dependent upon artificial intelligence and deep learning as a critical element in experimental and/or computational work. Featuring the Habana Gaudi training and first-generation Habana inference processors, Voyager encompasses a high-performance, low latency 400 gigabit-per-second interconnect from Arista. Voyager provides researchers with the ability to work on 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.

Voyager is an NSF-funded system, developed in collaboration with Supermicro and Intel's Habana Lab. It began a three-year testbed phase in early 2022 that will be followed up with a two-year allocation phase to the broader NSF community and user workshops. During the testbed phase, Voyager is available to select focused projects, as well as workshops and industry interactions. To request access to Voyager, please send a request to HPC Consulting at consult@sdsc.edu.





Triton Shared Computing Cluster

SDSC's Triton Shared Computing Cluster (TSCC) is UC San Diego's primary research HPC system. Its operations and user support are overseen by SDSC's Subhashini Sivagnanam. TSCC is foremost a "condo cluster" (researcher-purchased computing hardware) that provides access, colocation and management of a significant shared computing resource. It provides three kinds of compute nodes in the cluster: general computing nodes, GPU nodes and large memory nodes.

Recently, a UC San Diego team of bioengineers used TSCC to test a bioinformatics tool called SigProfilerMatrixGenerator, a new tool for classifying and visualizing large-scale mutational events—mutations that affect more than 50 DNA base pairs.

"Our new bioinformatics tool revolutionizes the way we visualize and explore structural variations (SV) and copy number variations (CNV) to decipher genomic anomalies in cancer development," said Ludmil Alexandrov, an associate professor of bioengineering and cellular and molecular medicine at UC San Diego. "Without access to SDSC's TSCC, we would not have been able to develop and test SigProfilerMatrixGenerator—we especially utilized the high-performance GPU and CPU computing on the cluster for our work."

Written in Python while also including an R wrapper, the team's tool accommodates two classification schemes for SVs and CNVs—allowing cancer specialists to analyze and visualize intricate mutational patterns across various cancer types. Specifically, SigProfilerMatrixGenerator is able to compound data from CV mutations into 48 different channels and SVs into 32 channels, to easily understand the pattern of genetic mutations and find their specific signatures. In terms of efficiency, SigProfilerMatrixGenerator has been developed to handle large datasets and can generate both a CNV and a SV count matrix for thousands of samples in a couple of seconds. Details of the tool are available in a recently published article entitled Visualizing and Exploring Patterns of Large Mutational Events with SigProfilerMatrixGenerator in the BMC Genomics journal (published Aug. 21, 2023).

Comet Retires

Comet, a cluster originally designed by Dell and SDSC to deliver 2.76 peak petaflops, is retiring. First operational in 2015, Comet was designed to transform advanced scientific computing by expanding access and capacity across domains. The most recent use of Comet was by the Center for Western Weather and Water Extremes (CW3E) team at Scripps Institution of Oceanography for their atmospheric rivers research and predictions. The supercomputer also worked to meet the needs of what is often called the "long tail" of science—the concept that numerous modest-sized computationally based research projects collectively represent volumes of research that can lead to scientific advances and discovery.

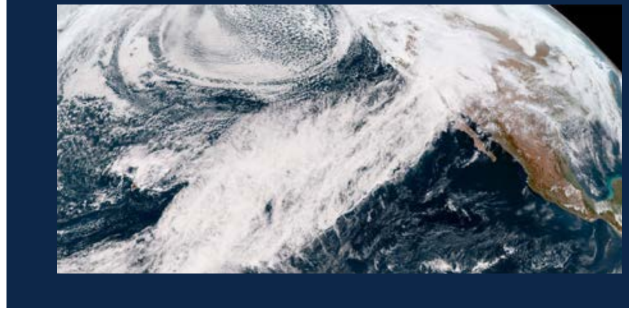
Improving prediction of atmospheric rivers

Amid a historic drought, a deluge of rain and relentless storms battered California this past winter, leading to severe flooding across the state and record snowpack in the Sierra Nevada. It's a prime example of "weather whiplash," which scientists say will become a more frequent occurrence as our climate continues to warm.

The weather phenomenon behind the bulk of this precipitation? Atmospheric rivers: the long, flowing regions in the sky that carry enormous amounts of water vapor, released over land in the form of rain and snow.

With the potential for high volumes of water to make landfall, preparedness is key. That's why, in the CW3E, atmospheric scientists like Luca Delle Monache, the center's deputy director, are using artificial intelligence—in the form of machine learning algorithms—to improve the prediction of atmospheric rivers. The future we've been warned about, he says, is already here.

"We are going to get more intense atmospheric rivers: the more impactful ones, the ones that create a lot of damage," said Delle Monache. "We're living in it now—these very intense storms, this record-breaking snow accumulation—may be caused by climate change."



When the atmosphere is warmer, it can hold more water vapor, which Delle Monache refers to as the "fuel" that powers an atmospheric river. And the more water vapor in the atmosphere, the more intense these storms can be. It's vital, then, that researchers at CW3E—a global leader in the study and forecasting of atmospheric rivers—are able to accurately predict Integrated Water Vapor Transport, or IVT, which is the key signature variable for determining the presence and intensity of these storms.

At CW3E, Delle Monache leads the Machine Learning Team, made up of atmospheric scientists and computer scientists who are improving the prediction of IVT by leveraging the power of massive amounts of weather data from the physics-based dynamical models used in forecasting. These models are imperfect, says Delle Monache, because the atmosphere is a chaotic system, meaning that even tiny errors in a forecast's initial conditions can grow quickly and significantly alter predictability.

The application of machine learning to the dynamical, physics-based model is a game changer," said Delle Monache, adding that this work has improved the prediction of IVT by as much as 20%. "It's an exciting time, where we're really making meaningful improvements and contributions."

By feeding the data from these models and observations into AI algorithms in what is called a "post-processing framework," Delle Monache and his team, using Comet, were able to improve their current predictions based on the errors the model made in the past.

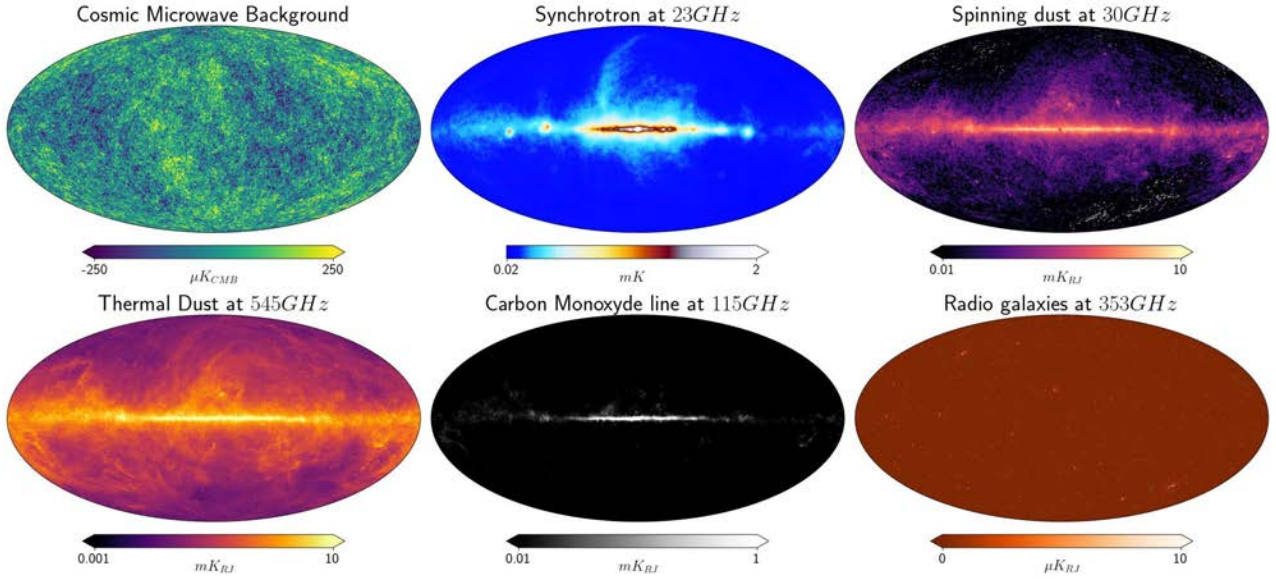
From Cosmology to Neuroscience, Data-Enabled Scientific Computing Group Highlights Two Key Projects

The Data-Enabled Scientific Computing (DESC) team at SDSC leads HPC and computational sciences innovation—serving and collaborating with thousands of national, UC-wide and UC San Diego researchers as well as industry partners. Providing full support and training to SDSC's user community, this year's DESC highlights include two key projects centered around neuroscience and cosmology.

The first is a research project that utilizes the NIH-funded NeuroIntegrative Connectivity (NIC) tool for EEG data processing to describe brain network analysis; the NIH- and NSF-funded Neuroscience Gateway (NSG) which freely and openly provides neuroscience software on supercomputers for neuroscientists; and the NSF-funded Open Science Chain (OSC) which utilizes consortium blockchain to maintain integrity of the provenance metadata for published datasets and enables independent verification of the dataset. Combining the OSC with the NSG platform, an efficient and integrated framework was created to automatically and securely preserve and verify the integrity of the artifacts used in research workflows for neuroscience software available on the NSG platform. The NIC tool was used as an example to demonstrate this framework where EEG and supercomputing related metadata from running the NIC tool via NSG, were stored on the OSC. Overall this contributed toward scientific reproducibility and FAIR principles. A paper by first author DESC's Cyberinfrastructure Solutions and Services Lead Subha Sivagnanam titled "Towards building a trustworthy pipeline integrating Neuroscience Gateway and Open Science Chain" was published in the journal database (vol. 2024: article ID baae023).

The second project, in collaboration with NERSC, leads the sky simulation effort in the CMB-S4 next generation cosmology experiment. CMB-S4 is an ambitious set of telescopes to be deployed simultaneously in Chile and at the South Pole to jointly probe the physics of the early universe—in particular hunting for traces of inflation. CMB-S4 has been recommended by the particle physics scientific community through the P5 report and is jointly funded by the NSF and DOE. Modeling and simulating the sky emission is an activity necessary to characterize the data and to support the decoupling of the primordial signal from emission from the Milky Way and other galaxies.

SDSC is also involved in sky-signal characterization and data delivery for the Simons Observatory, funded initially by the Simons Foundation and recently awarded a multimillion-dollar NSF award for its Advanced Simons Observatory expansion.



The figure shows pictures of some of the astrophysical models available in Python Sky Model (PySM). Each image is a full-sky map simulated with PySM of the intensity of each emission (color-coded based on the colorbar below each image) at a specific electromagnetic frequency. Credit: Andrea Zonca

FAIR Skies Ahead for Biomedical Data Project Looking to Benefit Research Community

SDSC, along with the GO FAIR Foundation, the National Center for Atmospheric Research, the Ronin Institute and other partners, has been conducting data landscaping work funded by the Frederick National Laboratory for Cancer Research, operated by Leidos Biomedical Research, Inc., on behalf of the National Institute of Allergy and Infectious Diseases (NIAID). SDSC's Research Data Services Director Christine Kirkpatrick, who leads the GO FAIR U.S. Office at SDSC, serves as PI for the new project.

The NIAID Data Landscaping and FAIRification project seeks to benefit biomedical researchers and the broader community to generate and analyze infectious, allergic and immunological data. Using the FAIR Principles as a guide, the project team—offering a broad background to ensure that metadata, a set of data that describes and gives information about other data, for biomedical research is findable, accessible, interoperable and reusable (FAIR)—provides guidance on approaches to enhance the quality of metadata within NIAID and NIH supported repositories and resources that harbor data and metadata.

Structured trainings and guidance support stakeholders, including components from the model pioneered by GO FAIR leveraging established M4M workshops and adopting FAIR Implementation Profiles (FIPs). This work is underpinned by interviews with stakeholders and an assessment to explore the relationship between FAIR resources and scientific impact. The initial period of the federally funded contract, which runs through Sept. 30, 2024, is valued at $1.3 million.

Highlights of the team's expertise include co-authoring the FAIR Guiding Principles, facilitating metadata for machines (M4M) workshops, developing the FAIR Implementation Profile approach, and contributing to improvements on data policy and metadata practices and standards.

"Our team is elated to be working with our NIAID project sponsors at the Office of Data Science and Emerging Technologies (ODSET) through Leidos Biomedical Research," remarked Kirkpatrick. "NIAID is renowned for its significant data resources and impactful scientific research. Having the chance to apply our collective expertise in research data management in support of the NIAID mission areas of infectious disease, allergy and immunology will be both impactful to the FAIR ecosystem, and meaningful work for our team. Further, I believe this work will become more common in the future as organizations begin to see data as a strategic asset, rather than focus on the cost of storing it."

The project follows alongside another key project in the Leidos Biomedical Research portfolio, the NIAID Data Ecosystem Discovery Portal, led by The Scripps Research Institute. The project team works hand-in-hand with the Scripps team to ensure repository improvements maximize the Discovery Portal's ability to search across the wide array of data assets produced by NIAID-funded research.

The project team includes co-authors of the 2016 FAIR Principles paper (Barend Mons and Erik Schultes), leaders in research data consortia, scholars in informatics, biomedical research and pioneers in FAIR training, interoperability practices and methodology for assessing scientific impact. Team members are Chris Erdmann, Doug Fils, John Graybeal, Nancy Hoebelheinrich, Kathryn Knight, Natalie Meyers, Bert Meerman, Barbara Magagna, Keith Maull and Matthew Mayernik. These experts are complemented by world-class systems integrators and project managers from SDSC: Alyssa Arce, Julie Christopher and Kevin Coakley.

Sustainable Scientific Software Division (S3D) Drives Innovation with Transformative Milestones in Democratizing Scientific Access

This year, the S3D underwent significant changes to enhance its mission of democratizing access to scientific resources through Science Gateways. These developments encompassed structural shifts, new personnel, awards and platforms—all aimed at expanding the division's impact on national cyberinfrastructure.

The SGX3 Center of Excellence for Science Gateways launched its inaugural Blueprint Factory activity at PEARC23, uniting cyberinfrastructure professionals and domain scientists to study the upcoming needs of AI-enabled research. Insights from 40 attendees highlighted the evolving landscape, emphasizing data handling and verification over additional compute resources. This activity will continue during the next year with additional participants from a variety of science domains, resulting in an analyst-style report describing the findings.

A substantial investment was made in creating the new Science Gateway platform, OneSciencePlace®. Reflecting a shift from computing-centric activities to prioritizing data, this platform is where "content meets computing." It integrates robust data management capabilities from the SeedMe Lab project with computing services like those offered by the Hubzero® platform. Notable projects, Quakeworx and CIPRES, are already lined up to take advantage of OneSciencePlace.

The Hubzero platform thrived with notable projects such as U.S. Pharmacopia's Continuous Manufacturing Knowledge Center and SDSU's HealthLINK project that focuses on promoting health equity and well-being for all people and communities. Moreover, Hubzero was chosen for the NSF's launch of the NAIRR Pilot portal—showcasing S3D and SGX3's involvement in the national AI initiative.

Sandra Gesing, from the University of Illinois' Chicago's Discovery Partners Institute, joined the group, bringing extensive experience in Science Gateways and research software engineering. Her role is a joint appointment with SDSC that involves intensifying involvement with SGX3 and with the U.S. Research Software Engineer Association as its inaugural executive director.

S3D's SDx group is fulfilling its goal of bringing professional software development, operations and sustainability to academic projects. SDx is staffed with a full complement of personnel with a wide variety of expertise, ranging from web development through low-level systems, databases, middleware, cybersecurity and user experience design. This breadth of experience allows SDx clients to have access to portions of individuals' time with the right expertise in just the right amounts needed for their projects. Such a staffing model is essential for academic projects with limited budgets that typically cannot afford full-time employees with all the skills they need. SDx has continued to support development of Science Gateway platforms and this year began work on the Pelican project at the OpenScienceGrid Consortium to add support for S3 storage in their data cache system.

S3D has had a robust year, serving the research community, shaping its future and actively contributing to the advancement of science gateways and cyberinfrastructure on a national scale.

INSIGHT: COMPLIANT, SECURE DATA

Sherlock Partners with BU's School of Public Health Biostatistics and Epidemiology Data Analytics Center to Satisfy Protected Data Research Needs

The Sherlock Division at SDSC has partnered with the Boston University (BU) School of Public Health Biostatistics and Epidemiology Data Analytics Center (BEDAC) to provide the requisite compliant and secure environment to protect the sensitive data for multiple investigators working in support of BEDAC's research with the Centers for Medicare & Medicaid Services (CMS).

Sherlock has supported a number of initiatives that involve CMS data and has worked cooperatively with partners to develop innovative solutions to ensure data is protected and secure so that researchers have the necessary services, tools, and infrastructure to manage and expand their research. For instance, Sherlock collaborated with CMS to build and deploy a FISMA-certified, high-performance data warehouse and analytics platform. This platform provided deep mining and analysis software tools to enable CMS's Medicaid Integrity Program to identify instances of Medicaid fraud, waste and abuse. Sherlock also delivered a HIPAA-compliant, cloud-based data warehouse and analytics platform to support the City of Hope's California Teachers Study. This research infrastructure incorporates data management and analytics solutions that have modernized and transformed the way in which the City of Hope protects and secures sensitive data and research.

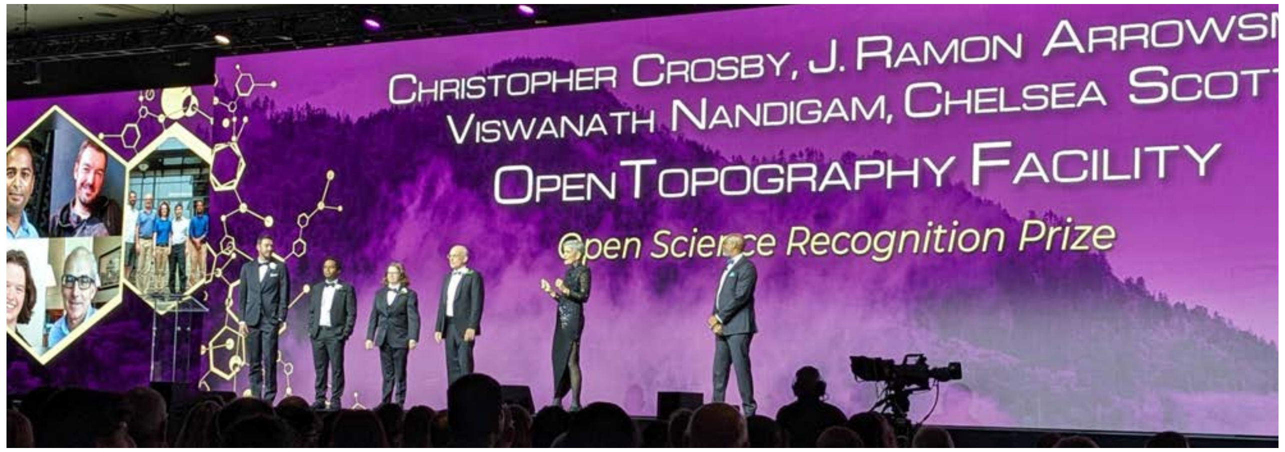
The addition of BEDAC to Sherlock's portfolio of partner academic institutions is well-positioned for success. BEDAC collaborates with investigators across all BU schools and departments, along with investigators at external institutions—in support of academic, government, foundation and industry sponsored research across various research disciplines. BEDAC has been at the forefront of analyzing insurance claims data, including commercial insurance as well as Medicaid and Medicare resources. BEDAC was among the first centers in the country to incorporate data from the Transformed Medicaid Statistical Information System (T-MSIS), a nationally representative Medicaid database, into translatable research.

The partnership between Sherlock and BEDAC draws upon Sherlock's knowledge and expertise protecting and securing data requiring the heightened compliance requirements outlined in FISMA and HIPAA. Together, Sherlock and BEDAC develop and deploy a novel FISMA-certified and HIPAA-compliant Cloud-based digital warehouse to support BEDAC's research with CMS. The team provides a cost-effective and user-friendly resource that meets the regulatory requirements and optimizes an environment for complex and high-impact Medicare and Medicaid data.

This resource allows for cross-collaboration while managing high priority data sources leveraging latest technologies to advance research. Sherlock is well-equipped and well-versed to seamlessly satisfy the depth and sensitive nature of BEDAC's varied research needs relating to CMS while providing the agility and scalability to accommodate the complexities of this project along with new research and projects.

Providing secure computing services in the cloud allows BEDAC staff to focus on direct support of research teams - instead of worrying about infrastructure and compliance, they can focus on mapping research problems to computational tools. This partnership supports the shift of focus by providing secure data storage and computing for the research teams.

INSIGHT: OPEN SCIENCE ACHIEVEMENT


Christopher Crosby, J. Ramon Arrowsm
Viswanath Nandigam, Chelsea Scott
OpenTopography Facility
Open Science Recognition Prize

OpenTopography Recognized for Excellence in Advancing Open Earth and Space Science

OpenTopography, a U.S. National Science Foundation (NSF)-funded data facility operated collaboratively between SDSC, EarthScope Consortium and Arizona State University (ASU), received the inaugural Open Science Recognition Prize from the American Geophysical Union (AGU) at its AGU23 conference in December 2023.

Open Science, according to UNESCO, is "an inclusive construct that combines various movements and practices aiming to make multilingual scientific knowledge openly available, accessible and reusable for everyone, to increase scientific collaborations and sharing of information for the benefits of science and society, and to open the processes of scientific knowledge creation, evaluation and communication to societal actors beyond the traditional scientific community."

The AGU recognition prize is awarded each year to a person or team for outstanding work in advancing Open Science related to Earth and space science and its impact globally. This year, OpenTopography principal investigators Viswanath Nandigam (SDSC), Christopher Crosby (EarthScope), Chelsea Scott and Ramon Arrowsmith (ASU) were recognized by the global Earth and space sciences community for their outstanding contributions in cyberinfrastructure, data management, training and outreach associated with open access high resolution topography.

Open Science Recognition Prize winners are chosen from nominations across the Earth and space science community for individuals or teams that advance Open Science through creation or use of open data, software and other open results. AGU, the world's largest Earth and space science association, annually recognizes a select number of individuals for its highest honors. Each recipient embodies the AGU community's shared vision of a thriving, sustainable and equitable future powered by discovery, innovation and action. According to AGU, these recipients have worked with integrity, respect and collaboration while creating deep engagement in education, diversity and outreach.

"This award is a tremendous recognition of the team who has built and maintained OpenTopography over the last 15 years. Credit is due to our advisory committee members past and present, as well as our funders, especially the NSF," said Nandigam. "And most importantly, it honors the trust of our partners and the vast community of users who rely on us to make available topographic data, data products and knowledge about those products and their applications open."

Since 2008 when it was founded, OpenTopography has been focused on facilitating efficient access to topography data, tools and resources to advance understanding of the Earth's surface, vegetation and built environment.

According to Arrowsmith, topography of the Earth's surface is a fundamental geomorphic and geophysical observable, marking the boundary across which the lithosphere, hydrosphere, biosphere and atmosphere interact. "These data are essential to the Earth sciences, and OpenTopography democratizes access to the data, tools and knowledge resources necessary to fully utilize the data in research and education," he said.

Crosby noted that it has been exceptionally rewarding to watch OpenTopography's user community grow and diversify as the system has become the most comprehensive source of topographic data on the internet. "OpenTopography's impact on research and education has been large, but the system is also regularly used by industry, governments and hobbyists from around the world," he said.

Center for Applied Internet Data Analysis

Founded in 1997 and based at SDSC, the Center for Applied Internet Data Analysis (CAIDA) investigates practical and theoretical aspects of the internet to provide macroscopic insights into internet infrastructure, behavior, usage and evolution; fosters a collaborative environment in which data can be acquired, analyzed and (as appropriate) shared; and improves the integrity of the field of internet science and informs science, technology and communications public policymakers.

CAIDA experts conduct network research and build research infrastructure to support large-scale data collection, curation and data distribution to the scientific research community. CAIDA receives funding through grants, gifts and sponsorship from governmental agencies, non-profit organizations and corporations.

Below are some of CAIDA's current projects.

ILANDS - Integrated Library for Advancing Network Data Science

This project is for enhancing the infrastructure to handle 100GB packet rates and projected routing table growth, including deploying enhanced storage and compute resources to support long-term use of the data.

Principal Investigators: kc claffy, David Clark
National Science Foundation (NSF CNS-2120399)

RABBITS - A measurement toolkit for Reproducible Assessment of BroadBand Internet Topology and Speed

This project aims to design, implement and deploy a measurement toolkit for Reproducible Assessment of BroadBand Internet Topology and Speed (RABBITS). This toolkit offers comprehensive and longitudinal datasets to facilitate scientific studies on topology and performance of evolving speed test infrastructure. The software included in RABBITS enables reproducible speed tests by supporting the use of consistent test parameters, even across different test platforms. The project provides rigorous internet measurement capabilities that can support broadband consumer protection efforts and identify opportunities to improve coverage in unserved/under-served communities.

Principal Investigator: Ka Pui Mok
National Science Foundation (NSF CNS-2323219)

STARNOVA - Unified Approach to Internet Performance Measurement

This project designs and implements the Sustainable Technology to Accelerate Research Network Operations Vulnerability Alerts (STARNOVA) platform. STARNOVA leverages unsolicited Internet traffic to provide early-warning indicators of new attacks targeted to cyberinfrastructure at SDSC. The newly developed infrastructure extends the visibility of the UC San Diego network telescope to monitor unsolicited traffic toward SDSC's equipment IP address spaces.

Principal Investigators: Ka Pui Mok, kc claffy, Fabian Bustamante
National Science Foundation (NSF OAC-2319959)

Cloud Bottlenecks - Detection and Analysis of Infrastructure Bottlenecks in a Cloud-Centric Internet

This project includes an effort to design measurement and analysis tools to reveal performance bottlenecks outside the cloud networks where the high cost of deployment and operations leads to infrastructure bottlenecks for cloud applications.

Principal Investigators: Ka Pui Mok, kc claffy, Alexander Marder
National Science Foundation (NSF CNS-2212241)

AVOID-5G - Automated Verification Of Internet Data-paths for 5G

The Automated Verification Of Internet Data-paths (AVOID) project focuses on providing unprecedented capability to tackle two high-risk attack vectors for 5G communications.

Principal Investigators: Alexander Marder, Erik Kline, Ka Pui Mok, kc claffy, Kyle Jamieson
National Science Foundation (NSF OAC-2326928)

MSRI-GMI3S - Designing a Global Measurement Infrastructure to Improve Internet Security

This project objective is to design and prototype a distributed, integrated infrastructure to measure the Internet, with the objective of improving Internet infrastructure security.

Principal Investigators: kc claffy, David Clark, Bradley Huffaker
National Science Foundation (NSF OAC-2131987)

QUINCE-NG - A Unified Approach to Internet Performance Measurement

The team develops and evaluates a fundamentally new approach to Internet performance measurement by conducting subjective assessments to measure the QoE of video streaming and conferencing applications in the wild and correlate these QoE measurements with Internet performance.

Principal Investigator: Ka Pui Mok
National Science Foundation (NSF CNS-2133452)

For questions and comments specific to these funding sources or activities, please email info@caida.org.

HPWREN Increases Network Performance and Reach to Support Sites across Southern California

The High Performance Wireless Research and Education Network (HPWREN) originated as a wide-area wireless communications research project nearly 25 years ago. Initially limited to San Diego County in support of internet data applications in the research and education domains, it evolved over time. A substantial collaboration with first responder teams emerged early in the project with informal discussions that led to agencies assisting with HPWREN's deployment of routers and wireless links on isolated mountaintops.

The initial first responder agencies with which HPWREN collaborated were the California Department of Forestry and Fire Protection (CDF), which is now CAL FIRE, via their Emergency Command Center, as well as the Regional Communications System staff at the San Diego Sheriff's Department. In retrospect, without this assistance atop these agencies' mountaintop locations, it would have been unlikely that HPWREN could have expanded to the scope it is today. Over time, collaborations also expanded to include San Diego Gas and Electric, ALERTCalifornia, as well as various research groups and educators at multiple academic institutions and organizations.

Examples of early connections range from biological field stations to astronomy observatories to tribal learning centers. These first research and education collaborations with HPWREN include Native American education centers, the San Diego State University Santa Margarita Ecological Reserve, the Mount Laguna and Palomar Observatories, the California Wolf Center and many earthquake as well as additional geophysical sensors.

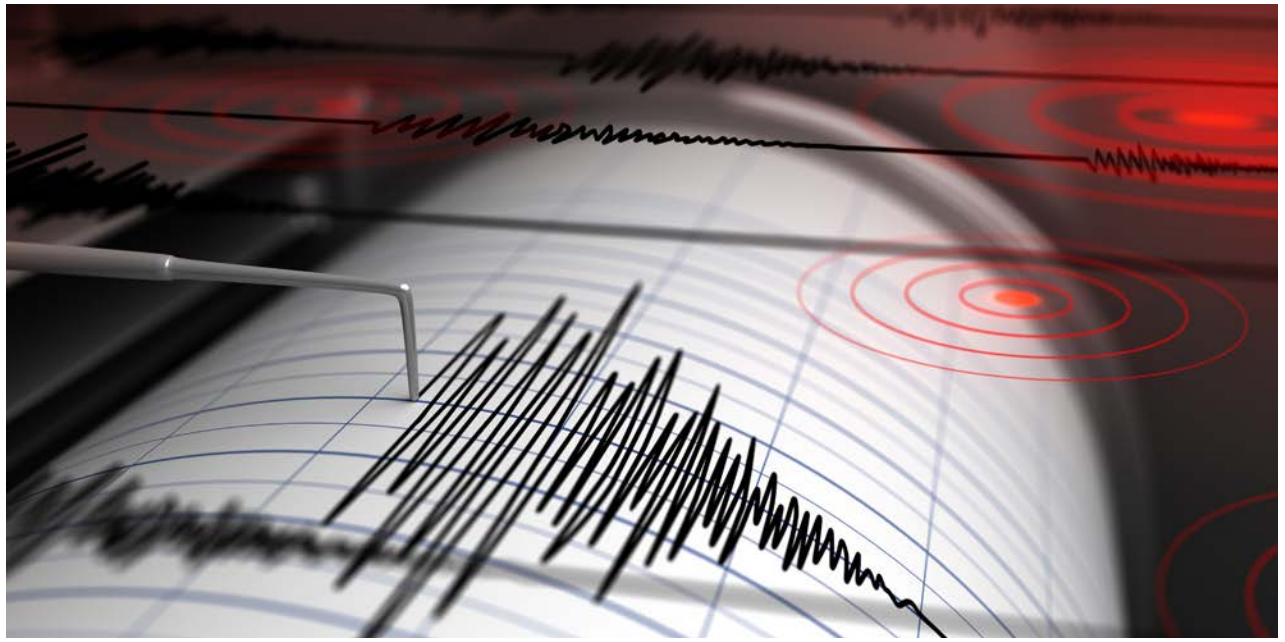
In addition to the communications infrastructure system for research and education, HPWREN provides wireless communications support for cameras and sensors, often in remote areas, through a radio-based wireless networking system. Many HPWREN-connected sites have been equipped with cameras and meteorological sensors, which allows first responders, researchers and educators to access for real-time and historical data. Most of the HPWREN-connected cameras continuously show 360-degree views in both color and monochrome.

"HPWREN produces open data—with the intent to make it broadly available in real-time," said HPWREN Co-founder Hans-Werner Braun.

Besides maintaining network stability, over the last several years the most important HPWREN objective has been to increase the network performance and reach to support more sites across southern California as well as to substantially increase the speed and redundancy of the backbone network that maintains the system.

"Significant progress has been made in network expansion into San Diego, Orange, Los Angeles, Santa Barbara, Imperial and Riverside counties, as we continue to expand reach and fill in gaps of coverage as defined by the fire agencies," Braun said.

More information is at https://hpwren.ucsd.edu/news/20231231/ and https://hpwren.ucsd.edu/



Shaking Up Earthquake Studies by Increasing Access to Data, Tools and Research Results

There is a rumbling in the seismic research community. Current studies about earthquake rupture forecasts or ERFs provide information about the probabilities of when earthquakes will occur, where they'll take place and how strong they will be. But advanced physics-based models that track fault system evolution backed up by observations are not accessible to a wide scientific community, which reduces the pace of research advancements and delays the benefits to society.

So scientists from the Southern (transitioning to Statewide) California Earthquake Center (SCEC) at the University of Southern California (USC), SDSC, the Scripps Institution of Oceanography (Scripps) at UC San Diego, and the University of Illinois Urbana-Champaign (UIUC) formed a team to address the problem. Their proposal to create a new science gateway for the community of researchers who study ERFs caught the attention of the U.S. National Science Foundation's Office of Advanced Cyberinfrastructure, which awarded the project a five-year $2.5 million grant.

The team for this collaborative grant award includes the project's Lead Principal Investigator (PI) from SCEC/USC, Yehuda Ben-Zion; SDSC/UC San Diego PI Amit Chourasia and Scripps/UC San Diego Co-PI Alice-Agnes Gabriel, and UIUC PI Ahmed Elbanna.

With goals to advance the science of ERFs, obtain new insights on earthquake physics, improve seismic hazard estimates and provide new opportunities for STEM education and engagement, the team is creating a cyberinfrastructure platform for the broad community to easily use computational tools and access data. Named Quakeworx, the science gateway provides access to state-of-the-art methods for physics-based simulations and data analysis. It hosts standardized pipelines for advanced simulations, facilitates data assimilation to inform and validates models on demand and enable machine-learning pattern recognition analyses on big datasets.

"Quakeworx will accelerate innovation in earthquake science by enabling generation of diverse outputs such as seismicity, ground motion, fault network configuration, strain rates and topography that can be used to validate model results, improve ERFs and discover new patterns," said Ben-Zion, director of SCEC and professor of Earth Sciences at USC Dornsife College of Letters, Arts and Sciences. "We anticipate the project to be transformational in research of earthquakes and faults, while contributing to modernizing education and facilitating translation of results to society at large."

Ben-Zion noted that the project enables large-scale simulations of coupled evolution of earthquakes and faults for the first time. "Quakeworx will curate and seed many leading simulation tools and data products that leverage existing SCEC's community data models describing fault and rock properties as simulation input," he said.

According to Chourasia, associate director for the S3 Division at SDSC, who leads the gateway cyberinfrastructure effort, Quakeworx builds upon and improves existing NSF investments such as Hubzero®, OneSciencePlace®, Tapis and SeedMeLab. It provides a platform to curate state-of-the-art simulators and machine learning applications in seismology that could be easily executed on a variety of local and remote computation resources such as those available at SDSC and from NSF's ACCESS program. Overall, Quakeworx delivers an innovative, service-oriented and easy-to-use cyberinfrastructure to a large user community.

"The developed Quakeworx gateway will enable the community to contribute, curate and share tools for simulations, data analysis and visualization, data products and knowledge base on an open platform. It will provide a foundation that provides build-once-and-reuse-for-all that eliminates barriers to lack of expertise and lack of access to compute resources," said Chourasia, who is also the director of Hubzero, adding that the Quakeworx framework supports extensibility through a pluggable architecture and composability with other systems and include first-class implementation that makes content FAIR (Findable, Accessible, Interoperable, Reusable).

Gabriel, associate professor at the Institute of Geophysics and Planetary Physics at Scripps, and Elbanna, associate professor and the Donald Biggar Willett Faculty Fellow at UIUC, lead the development and integration of several physics-based Next Generation Earthquake Simulators, including the current NSF Leadership-Class Computing Facility application SeisSol, utilizing high-performance computing for the project. They noted that the reduction in barriers to the use of computational methods in earthquake research will be achievable with the Quakeworx gateway that is designed to simplify access to state-of-the-art research software, compute and storage resources.

They also lead together with Ben-Zion the development and integration of machine learning apps for the project, and they said further that Quakeworx enables rapid immersion of students, post-doctoral fellows and early career scientists with state-of-the-art simulation and analysis tools. This democratization of access aids in modernizing education and training in earthquake physics, computational modeling, software engineering and related fields.

The project team engages and mentors four postdoctoral fellows, six graduate students, 15 undergraduates and 10 professionals. Ben-Zion explained that SCEC's role in the project is to coordinate the science needed to develop and extract information from the next generation earthquake simulations, while SDSC's role is to develop the computational gateway needed to support the next generation simulations and data analyses.

Broader impacts of the project include:

- Contributing to the development of a more resilient society to earthquakes, including working to provide critical information for emergency services and design of rapid follow-up scientific studies after large earthquakes;
- Sharing of key results and services with the broad earthquake community including academic institutions, federal and state agencies, and the private sector, as well as translation of advanced computational research into publicly available scientific information; and
- Enabling workforce development and community engagement via mentoring, training workshops and dissemination of results through conferences and scholarly articles.

According to the team, the project contributes to the national interest by supporting a more resilient society with improved capabilities for forecasting earthquakes and assessing seismic hazard, which are foundational for mitigation of seismic risk, contributing to STEM workforce development and facilitating translation of results to the society at large.

Support for this project is from the NSF (grant awards nos. 2311206, 2311207 and 2311208).

A firefighter supervises a controlled burn at Marine Corps Base Camp Pendleton, California. Credit: Sgt. Jake McClung, U.S. Marine Corps

WIFIRE Lab Forms New Partnership with U.S. Department of Homeland Security

For the past 10 years, the WIFIRE team at SDSC has been focused on meeting the growing needs of hazard monitoring, mitigation and response. Most recently, the team partnered with the U.S. Department of Homeland Security (DHS) to integrate edge computing—a strategy emphasizing data collection and analysis at the site of or geographically near data sources. To showcase the technology's impact in hazardous settings such as wildfire response and mitigation, the team completed a concept demonstration for a prescribed burn scenario.

"The WIFIRE infrastructure turns data into a utility for wildland fire scientists and the managers who make decisions based on insights from the data," WIFIRE Lab Founding Director and Principal Investigator Ilkay Altintas said. "Our WIFIRE Edge platform supports both wildfire response and mitigation scenarios using edge technology, and the concept demo highlights the integrating of these tools for both operational use and research in wildland fires."

Led by Altintas, SDSC's chief data science officer and director of CICORE, the WIFIRE team includes fire scientists, machine learning experts and data science researchers collaborating to help the fire management community combat wildfires and prevent future occurrences. Edge computing provides the team with even more powerful tools.

"Edge technology allows researchers to garner data that will lead to a deeper understanding of wildland fire," said WIFIRE Lab Director of Product Management Shweta Purawat. "Edge computing reduces networking requirements by minimizing the need for data to be transferred and processed at a remote data center."

Prescribed burns are conducted in order to limit an uncontrollable wildfire in the future by reducing hazardous fuel loads that lead to increased risk.

WIFIRE Edge platform integrates live ground sensor data including live ignitor positions, various environmental measurements and air quality parameters. The platform uses a containerized edge service developed by the WIFIRE Lab team that runs on edge compute devices, enabling real-time Fine Dead Fuel Moisture (FDFM) data for prescribed burns. Sensor units developed by Red Line Safety, Inc. are used for this purpose. In the recent demo, the WIFIRE team pulled in critical pieces of information to their existing BurnPro3D platform, using sensors to monitor real-time conditions and calculate key determinants of fire behavior to lay the groundwork for modeling how prescribed burns will evolve given current weather conditions and actual ignition patterns. BurnPro3D is developed with U.S. National Science Foundation support in collaboration with many partners, including LANL, U.S. Geological Survey, U.S. Forest Service and USC.

"We imagine a world where, in the future, technologists, scientists and fire managers will collaborate on unique solutions tailored to keep our communities safe and make the best decisions for mitigation and response to fires," Altintas said. "This new collaboration with DHS allows us to further these efforts and enables us to demonstrate the pathways and scenarios for utilization of edge technologies in wildland fire management."

WIFIRE was funded by the NSF (grant no. 1331615) under CI, Information Technology Research and SEES Hazards programs (NSF grant nos. 2040676 and 2134904) Convergence Accelerator program, Department of Homeland Security (grant no. 311098-00001), NOAA, NASA, San Diego Gas & Electric, CalOES, Intervalien Foundation, and additional industry and fire agency partners.

Firing Up Fire Detection Efforts with Deep Learning Models

WIFIRE lab researchers have developed new deep learning models to continue improving efforts for early wildfire detection. Recent research experiments integrated detailed on-the-ground, real-time camera footage, satellite-based fire detections and weather data to provide a multimodal approach to the early detection of wildfires.

Led by SDSC's Lead for Data Analytics Mai H. Nguyen and UC San Diego Computer Science & Engineering Professor Garrison W. Cottrell, the team published their work in a paper titled Multimodal Wildland Fire Smoke Detection in Remote Sensing, MDPI.

The paper discusses the team's utilization of deep learning models, which are artificial intelligence (AI) models that use multiple processing layers to learn representations of data at increasingly complex levels of abstraction. Using these representations, the model can detect patterns that can be used to make predictions. The models specifically presented in this research include the SmokeyNet baseline model, SmokeyNet Ensemble and Multimodal SmokeyNet extension.

The baseline SmokeyNet is a spatiotemporal model consisting of three different deep learning models: Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM) and Vision Transformer (ViT). The SmokeyNet Ensemble, designed to merge the SmokeyNet baseline model's image-based smoke predictions with weather data and satellite-based fire predictions, computes a weighted average of fire predictions obtained from processing these three data sources. Meanwhile, the Multimodal SmokeyNet, an extension of the SmokeyNet baseline model, directly integrates weather data with camera images.

"We use the Fire Ignition images Library (FIgLib) dataset, which consists of about 20,000 images across geographically distinct terrains throughout southern California, as the source for the camera data," said Jaspreet Bhamra, first author on the paper and former machine learning intern at SDSC. "These images were processed by SmokeyNet to detect wildfire smoke. We also use information regarding weather near the location of the images from three weather stations closely neighboring the cameras."

For the SmokeyNet Ensemble, fire detection data based on satellite images from the Geostationary Operational Environmental Satellite (GOES) system was also used; this data was designed to detect and characterize fire from biomass burning. Bhamra said that their results showed no notable advantage from the SmokeyNet Ensemble over the SmokeyNet baseline model. This indicated that the weather data and the GOES satellite data both served as weak signals. In contrast, the Multimodal SmokeyNet showed a distinct improvement concerning accuracy, F1, and time-to-detect in comparison to the SmokeyNet baseline model.

"A 13.6 percent improvement in time to detect initial smoke was reported, and F1 mean improved by 1.10 while standard deviation decreased by 0.32 on average for smoke detection performance, allowing us to conclude that the Multimodal SmokeyNet model was the most efficient and stable of the three models tested," Bhamra said.

Next steps for the team's work include expanding this work to additional fires and making use of unlabeled data to further improve performance.

"Specifically, we plan to analyze data from different geographical locations and camera types to extend the generality of our approach," Nguyen said. "The issue of false positives, including low clouds, is also being examined in order to improve detection performance."

Nguyen said that the team is also exploring methods to optimize the model's compute and memory resource requirements for effective real-time smoke detection to assist with the battle again wildfires and related destruction.

This work was funded by the NSF (grant no. 1541349) and San Diego Gas & Electric. Computational resources were provided by NSF (grant nos. 2100237, 2120019, 1826967 and 2112167).

AI Leadership with Chatting GPT

Artificial Intelligence (AI) systems are quickly emerging on the computing and data landscape. In particular, Large Language Models (LLMs)—AI systems trained on massive amounts of text—have reached a surprising level of capability, with the recent iterations able to write essays, poems, and computer code, and score near the 90th percentile on standardized tests such as the LSAT and the Math SAT. The most popular interface to this technology, ChatGPT, made the power of LLMs readily-available to the general public for the first time, and in doing so became the fastest-growing consumer application in history. It is clear that ChatGPT and other LLMs are having major impacts on how we work, learn, and live—and there is a sense that we have only seen the tip of the iceberg in terms of what these technologies can do.

In response to this, during spring 2023 SDSC in partnership with the Halicioğlu Data Science Institute (HDSI) hosted a series of talks and panels, targeted to the campus community and open to the general public. UC San Diego experts discussed ChatGPT and other generative artificial intelligence addressing what it is, how it works, what its ethical implications are, and what impacts it will have on fields such as medicine, business and education.

Speakers included Frank Würthwein, SDSC director and UC San Diego physics professor; Justin Eldridge, teaching professor at HDSI; Jingbo Shang, professor of computer science and engineering; David Danks, philosophy professor; Tiffany Amariuta-Bartell, professor HDSI/School of Medicine; Dr. Chris Longhurst, chief medical officer UC San Diego Health; Vincent Nijs, professor and associate dean of Academic Program, Rady School of Management; Robert Twomey, assistant professor, Johnny Carson Center for Emerging Media Arts at University of Nebraska-Lincoln; Mikhail Belkin, professor HDSI; Zhiting Hu, professor HDSI; Dheeraj Mekala, PhD student HDSI; Zihan Wang, PhD student HDSI; Stuart Geiger, professor HDSI/Communication; Tricia Bertram-Gallant, director of Academic Integrity Office; Shannon Ellis, teaching professor CogSci; Leo Porter, professor CSE.

The Chatting GPT event, in collaboration with UCTV, tracked 750 participants—590 unique online viewers and 160 onsite attendees—making the event SDSC's largest on record (aside from the annual Supercomputing Conference). Participation spanned the globe—India, UK, Australia, Canada, France, UAE, Mexico, Israel, Greece, Hong Kong, Saudi Arabia, Kosovo and Peru. SDSC's Programs and Events Manager Susan Rathbun co-led efforts for the event with HDSI's Justin Eldridge.

Experts Gather to Discuss Using Data and AI to Improve Responses to Mass Casualty Events

Late last fall, computer science and AI researchers, school safety managers, 911 program directors, service providers and policymakers gathered for the Predicting Mass Casualty Events from 911 Data Workshop at SDSC. The meeting focused on how best to utilize new types of 911 data streams for early detection of mass casualty events.

Led by SDSC's Director of Spatial Information Systems Laboratory and Member of the CICORE Division Ilya Zavlavsky, the workshop was facilitated by Don Reich from Public Safety Network Americas (PSNA) and Jon Whirledge from INdigital, who brought their extensive connections with state agencies and commercial companies enabling the 911 call infrastructure across the U.S.

"We discussed technical and organizational challenges and opportunities of using 911 data streams with advanced AI models to improve response to mass casualty events, especially on school campuses," Zaslavsky said. "By finding patterns across multiple calls coming from a geographic location, and utilizing Next Generation 911 (NG911) standards, we lay the groundwork of a novel analytical and predictive framework for a better-designed, nationwide prototype early alert for 911 responders."

The two-day event included 25 participants who discussed specific needs for improved 911 data models for real-time event detection and shared perspectives on robust emergency notification and response. The workshop was a crucial step in fostering multi-disciplinary collaboration among government, academic, industry and non-profit sectors to explore innovative uses of 911 data in public safety.

One of the event's sessions included a review of analysis opportunities provided by NG911 compared to legacy analog-based 911 data, and early experiences with AI modeling of the 911 data streams. "We were pleased to share our current project with the community," said Reich of PSNA, which provides 911 data analytics for public safety answering points throughout the U.S.

Another session at the event was a panel on school safety, where speakers articulated common challenges and insights learned from school shooting events. The panel was led by Tom Wheeler, Office of the Indiana Attorney General, and former Assistant Attorney General for Civil Rights, United States Department of Justice, who discussed School Safety: Lessons Learned with attendees. He was joined by speakers representing San Diego County District Attorney's Office, San Diego County Office of Education, and San Diego Psychiatric Emergency Response Team. Through a series of examples and case studies, the panel examined the vital role of school resource officers, the importance of timely data sharing and collaboration between various entities involved in school safety, and the role of new technical approaches in improving situational awareness and enabling rapid response to emergencies.

Discussion sessions at the event focused on data acquisition and data sharing for 911 data, predictive modeling, ethical and privacy concerns related to emergency notification, the critical public safety infrastructure, and the importance of partnerships for public safety efforts.

"Our primary take-away from the event was a better understanding of the need for a collective effort to overcome public safety challenges plaguing schools nationwide and technical issues involved in the overhaul of the legacy 911 system to make it more effective and efficient," Zaslavksy said. "As the next steps, we will build on the collaborations started at the workshop to advance predictive modeling and notification prototypes that have the potential to be widely deployed by our government and industrial partners."

Funding for this conference was provided by the U.S. National Science Foundation (grant no. 2330460).

SDSC Joins GA, Hewlett Packard and Sapientai to Support High-Priority Fusion Research

Late last summer, the U.S. DOE announced its selection of a multi-institutional team of data scientists from General Atomics (GA), SDSC, Hewlett Packard Enterprise (HPE) and Sapientai to develop a Fusion Data Platform (FDP) for advancing high-priority fusion research. In support of this effort the DOE awarded the team a three-year $7.4 million grant.

Led by GA, the FDP initially will be deployed at SDSC. Once completed, the FDP will be made available to the scientific community to provide access to high-quality fusion data for the efficient creation of reproducible artificial intelligence (AI)/machine learning (ML) models to support the design and operation of a broad range of fusion pilot plants (FPP) designs and plasma configurations within a decadal timescale.

A suite of AI/ML modeling capabilities developed by Sapientai and UC San Diego Computer Science and Engineering Faculty Rose Yu and Sicun Gao will be integrated with the platform, allowing it to serve as a powerful data and analysis tool that meets the growing needs of the fusion science community.

“Creating a robust AI/ML platform with very large curated datasets and efficient processing tools will be transformational for fusion energy,” said Brian Sammuli, head of the Fusion Data Science Center at GA and principal investigator. “By advancing AI/ML research in fusion, we will be able to rapidly address many of the remaining challenges in fusion science and reactor development. We look forward to leading this team to provide an outstanding platform for the scientific community to advance fusion research and support the deployment of the first generation of fusion energy power plants.”

According to Raffi Nazikian, senior director and leader of the ITER Research Hub at GA, a key mission of the FDP is to accelerate AI/ML research by expanding access to high-quality fusion data and the tools needed to process the data at scale.

“The FDP will include experimental and simulated data in an integrated platform. We are talking many petabytes of data that will be easily accessible on the platform,” said Nazikian. “The success of the FDP will be measured by how well we serve the needs of the fusion and broader data science community, including students and researchers from universities, national laboratories and industry.”

SDSC Director Frank Würthwein said that the FDP is an important step toward harnessing the power of fusion data to advance the development of fusion energy.

“GA and SDSC have a long history dating back almost 40 years, and this is the beginning of a new chapter in our cooperation to advance fusion energy science and education,” Würthwein noted.

Paolo Faraboschi, HPE fellow and AI Research Lab director at Hewlett Packard Labs, said that his team is excited to help build a powerful data platform for fusion. “Among the FDP unique capabilities will be the ability for users to access, understand and leverage prior data and AI pipelines to advance their research and build reproducible, certifiable AI/ML models. We look forward to working with the scientific community on the FDP to help realize the decadal vision for fusion energy development.”

SUPPORTING DATA-INFORMED FPP DESIGNS

To achieve fusion conditions relevant for energy production, an FPP must sustain plasmas at temperatures exceeding 100 million degrees Celsius—approximately 10 times the temperature at the center of the sun. In magnetic confinement fusion, plasmas are controlled using powerful electromagnets that shape and confine the superheated gas. At such extreme temperatures, the plasmas may exhibit instabilities that may cause them to momentarily breach the magnetic fields and interact with the inner walls of the fusion machine, which could decrease efficiency or even cause damage. Successfully designing FPPs that account for these and other types of instabilities requires robust data sets to model and predict plasma behaviors across designs.

The FDP will leverage GA’s scaleable, fusion-specific data processing tool, TokSearch, to process and curate the data sets at the required scale. The team will also draw from HPE’s Common Metadata Framework to create reproducible workflows that include metadata tracking, source code integration, and data version control. A publishing portal will be incorporated into the system to facilitate search and discovery of these curated datasets. A suite of AI/ML modeling capabilities developed by Sapientai and UC San Diego will be integrated with the platform, allowing it to serve as a powerful data and analysis tool that meets the growing needs of the fusion science community.

SDSC Among Group Awarded $25 Million NSF Grant for High Energy Physics Research

An international cohort of the Institute for Research and Innovation for Software High Energy Physics (IRIS-HEP), a software institute, headquartered at the Princeton Institute for Computational Science & Engineering, received some welcome news recently when the U.S. National Science Foundation’s Office of Advanced Cyberinfrastructure and Physics Division, awarded it $25 million for another round of research funding. Included among the collaborating institutions was SDSC, led by Professor of Physics Frank Würthwein, an expert in distributed high throughput computing and experimental particle physics.

IRIS-HEP provides key services for distributed high throughput computing (dHTC) through the OSG Consortium, an integrated open-source software stack, and a fabric of services for researchers and research organizations to accelerate their research via dHTC.

“Within its scope, IRIS-HEP provides the domain-specific contribution to the OSG Consortium for the Large Hadron Collider (LHC) community, R&D on new algorithms and the data infrastructure needed to support the exabyte of data expected to be produced annually by the LHC in the 2030s. Researchers at UC San Diego play leadership roles on projects in all three of these areas of IRIS-HEP,” said Würthwein, who also directs the OSG Consortium.

“The highly demanding computing and data requirements expected by the next LHC phase such as storing, transferring and processing half an Exabyte per year propose a challenge that is far from being met by the estimated performance gains from future hardware enhancements. IRIS-HEP’s goal is to close this gap by making software algorithms more efficient and, with the help of OSG, implementing innovative cyberinfrastructure techniques for a better use of the available resources,” said SDSC Scientific Software Developer and Researcher Diego Davila.

According to Saul Gonzalez, director of the U.S. National Science Foundation’s Physics Division, the amount of scientific data expected to be produced by the High-Luminosity LHC is staggering. “The innovative software and data analysis tools being developed by IRIS-HEP will allow researchers to find those ‘needles’ of discovery that unlock new understanding of our universe—and which would otherwise remain hidden in a haystack of data.”

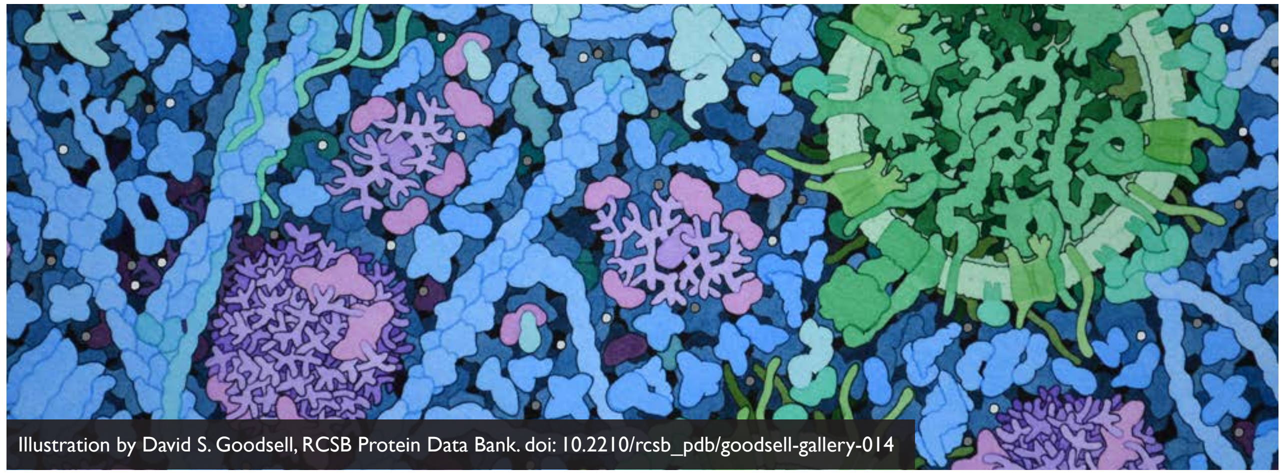


Illustration by David S. Goodsell, RCSB Protein Data Bank. doi: 10.2210/rcsb_pdb/goodsell-gallery-014

Protein Data Bank Remains a Key Resource for Scientific Discovery

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. In its 53rd year of continuous operations, the PDB has started a new five-year cycle with renewed funding from the U.S. National Science Foundation, the National Institutes of Health, and the Department of Energy, taking the PDB presence at SDSC to 30 years of consecutive funding.

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) as a collaboration between Rutgers and UC San Diego/SDSC. Since then, the RCSB PDB has operated the U.S. data center for the global PDB archive, making 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, has 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,” he said.

Approximately 60,000 structural biologists (depositors) working on every inhabited continent have contributed data to the archive. This information is used by millions of PDB Data Consumers (basic and applied researchers, trainees, educators and students) based in more than 200 UN-recognized sovereign countries and territories. The archive has been designated by the Global Biodata Coalition as a Global Core Biodata Resource, and CoreTrustSeal-certified.

From 2019 to 2023, more than 67,000 atomic coordinate files and related experimental data files were publicly released. During the same five-year period, PDB data files were accessed more than 10.7 billion times from wwPDB data centers around the world. During 2023 alone, PDB data were accessed from RCSB PDB more than 2.6 billion times. Additionally, 476 trusted external information resources repackage and distribute PDB data for the global scientific community. PDB data are also maintained as standalone copies of the archive inside for-profit company firewalls.

Last year, PDB announced that 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 agreed to contribute a high-performance server hosted at the SingAREN Open Exchange, located at Equinix SG3. SingAREN agreed to provide the high-speed international connectivity for the server in the region. Deploying the cache server at SingAREN provides researchers in Asia—particularly those working in the fields of genomics, climate science and materials science—with faster and more efficient access to data, enabling quicker and more efficient cutting-edge research and discoveries. Over the next two years, the two organizations will actively work together in support of the Open Science Data Federation and the RCSB PDB.

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

Images: Tetsuro Matsuzawa; CARTA; NASA

SDSC and CARTA Maintain Valuable Partnership

The Center for Academic Research and Training in Anthropogeny (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?”

Before COVID, CARTA’s events averaged 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 remained extremely successful, with hundreds of 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.

“CARTA’s long-term partnership with the San Diego Supercomputer Center has been extremely valuable,” said 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.”

New Award Funds SDSC to Conduct HPC Training for Domain Scientists

While successful researchers have the ability to find solutions to complex problems, often the implementation of their computational methods requires a machine that is bigger and more powerful than their desktop computer. Chemists, environmental scientists, geophysicists, biologists and others often find themselves in need of a supercomputer. These domain scientists and applied mathematicians typically work alongside a computational scientist to implement and run their workloads; however, thanks to an award from the U.S. National Science Foundation (NSF) to SDSC, a training program has been developed to better assist with learning the ins and outs of supercomputing.

"We are pleased to offer COMPrehensive Learning for end-users to Effectively utilize Cyberinfrastructure, or COMPLECS, training for our SDSC resource users beginning spring 2024," said Bob Sinkovits, director of training at SDSC and principal investigator (PI) for the new three-year $500,000 award. "Nicole Wolter and Marty Kandes, also at SDSC, are co-PIs on the project and we have been working on preliminary plans for this program since our summer institute, which will serve as a model for COMPLECS."

The COMPLECS training program consists of three layers—beginning with foundational knowledge, such as parallel computing concepts and intermediate Linux, that serves as a base for learning other essential and specialized skills depending on the users' needs. The program will host multi-day in-person workshops and webinars as well as a self-paced online study option.

Additional SDSC community members involved with the COMPLECS effort are Advanced Computing Training Lead Mary Thomas and Andreas Goetz, who is the director of the center's Computational Chemistry Laboratory.

"One of our goals with the program is to recruit participants from underrepresented groups and domains that haven't traditionally used supercomputers in conjunction with our work on the ACCESS-CI project," Thomas said. "We are just thrilled to have this opportunity and grateful to the NSF for funding our work in this arena."

The program is funded by the NSF Office of Advanced Cyberinfrastructure (award no. 2320934).

SDSC Leaders Collaborate with Schmidt AI Fellow on Data Workflow and Knowledge Management

Last spring, UC San Diego announced that 10 postdoctoral scholars would be applying artificial intelligence (AI) methods to their research in a range of fields with support from the Eric and Wendy Schmidt AI in Science Postdoctoral Fellowship, a program of Schmidt Futures. One of those scholars is Jessica Kendall-Bar, who recently joined the Scripps Institution of Oceanography Center for Marine Biotechnology and Biomedicine.

As a Schmidt AI Postdoctoral Fellow, Kendall-Bar is co-mentored by Ilkay Altintas, chief data science officer and director of CICORE at SDSC. Altintas is also the associate director for education in the fellowship program, working with Program Director Tara Javidi, professor of electrical and computer engineering at the Jacobs School of Engineering, to help shape the outcomes of the program through education and training objectives.

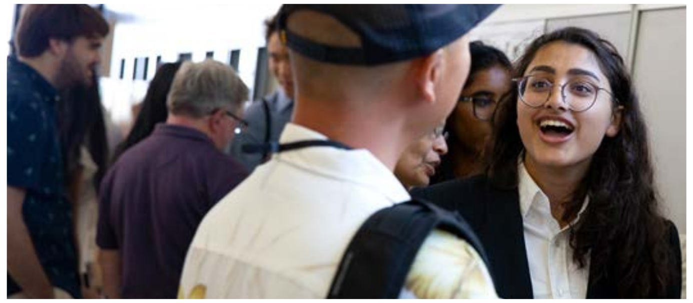
"We are excited to have Jessica Kendall-Bar join us as a fellow," Altintas said. "While eager to assist her with her goals of learning new workflow and knowledge management techniques as well as 3D visualization, we are thrilled to learn from her about her work on the physiology of wild seals and improved science communication."

Since 2012, Kendall-Bar has been studying and conducting research within the University of California system at UC Berkeley, UC Santa Cruz and now UC San Diego. For her doctorate in ecology and evolutionary Biology at UC Santa Cruz working with Terrie Williams and Dan Costa, she built a portable system for recording sleep in wild northern elephant seals, discovering that they sleep deep beneath the ocean's surface. She explored her data using custom data workflow and visualizations that were featured in the paper published in Science as well as news outlets like The New York Times and The Atlantic. As an award winning science communicator, Kendall-Bar combines analysis and immersive visualization to find patterns and communicate those to inform international decision-makers on topics ranging from marine mammal conservation to coral reef restoration.

"I'm grateful for the opportunity to work with Dr. Altintas and her team of data experts to build new tools for scientists," Kendall-Bar said. "In order to leverage AI for data analysis, we need to create computational workflows to quickly visualize and analyze large timeseries. These visual analyses can help inform and communicate urgent messages about the state of ecosystems and our climate."

UC San Diego was one of nine top universities worldwide selected to partner on the $148 million initiative to support postdoctoral fellows as they learn and apply AI methods to their research. In addition to Altintas, SDSC Director Frank Würthwein is one of the STEM champions in the program along with other UC San Diego faculty.

The fellowship was created by Schmidt Futures, a philanthropic initiative of Eric and Wendy Schmidt, to change how science is done by accelerating the incorporation of AI techniques into the natural sciences, engineering and mathematical science, providing access to AI tools and training to the sharpest minds on the frontlines of scientific innovation.



SDSC Programs

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.

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.

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.

RESEARCH EXPERIENCE FOR HIGH SCHOOL STUDENTS

SDSC's Research Experience for High School Students (REHS) program, which celebrated its 14th year in 2023, was developed to help increase awareness of computational science, technical writing and related fields of research among students in the greater San Diego area. The eight-week program paired SDSC mentors with high school students to help them gain 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 their peers, other 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 invited to apply to UC San Diego's Mentor Assistance Program (MAP), a campus-wide initiative designed to engage students in a mentoring relationship with an expert 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 biology, chemistry, aerospace engineering, network architectures, pharmaceutical sciences, physics, social sciences and more.

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 again 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.

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 2023 Summer Institute continued its two-part structure in response to the needs of a diverse user base and the increasingly diverse suite of resources and services that they need to utilize—often referred to as cyberinfrastructure. The workshop also served to address the needs of non-programmers who need to acquire specialized skills to effectively use advanced cyberinfrastructure.

HPC TRAINING WEBINARS AND WORKSHOPS

A vast array of HPC training opportunities were offered over the past year at SDSC. Multiple events focused on familiarizing researchers with HPC systems such as Expanse, Voyager and the Triton Shared Computing Cluster. The Center's HPC programs 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.

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 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 to create new CCRs.

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 the Ivory Coast to Bangladesh. A subset of students pays for a certificate of completion.

CYBERINFRASTRUCTURE-ENABLED MACHINE LEARNING (CIML) SUMMER INSTITUTE

The CIML Summer Institute introduces machine learning (ML) concepts to researchers, developers and educators, who need techniques and methods to migrate their ML applications from smaller, locally run resources, such as laptops and workstations, to large-scale HPC systems, such as SDSC's Expanse supercomputer. In 2023, participants had the opportunity to accelerate their learning process through highly interactive classes with hands-on tutorials using Expanse.

STUDENT CLUSTER COMPETITION

The Student Cluster Competition (SCC) is an event featured each year at the annual International Conference for High Performance Computing, Networking, Storage, and Analysis (SC). Implemented in 2007, the event immerses undergraduate and high school students in high-performance computing. Teams consist of a mentor plus six students who design and build a small cluster with hardware and software vendor partners. They learn designated scientific applications and apply optimization techniques for their chosen architectures. SCC teams compete against teams from around the world, in a non-stop 48-hour challenge to complete a real-world scientific workload, while keeping the cluster up and running—all to demonstrate to the judges their HPC skills and knowledge. Acceptance to competition is stiff and requires intense preparation and skill development. At SC23, held in Denver, Colorado, a team of UC San Diego undergraduate students won the MLPerf Contest and placed third overall in the annual competition. The students from SDSC and Jacobs School of Engineering called their team Triton LLC (Last Level Cache). They were among a total of 11 in-person teams from around the world selected to compete.

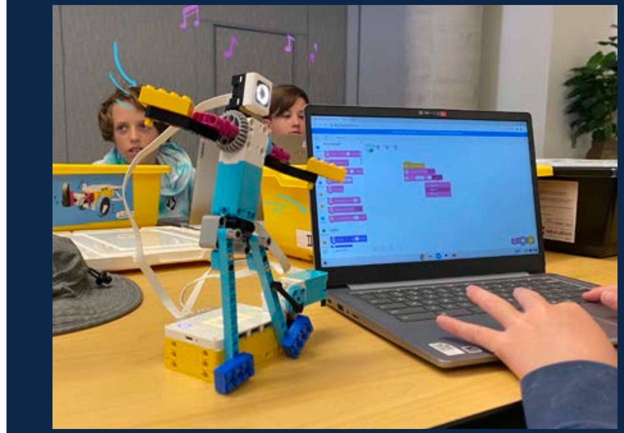
CORE FELLOWS

The Convergence Research (CORE) Institute, funded by the NSF Convergence Accelerator, is designed to catalyze an impact network of researchers, practitioners and industry and public policy professionals committed to collaboratively engaging in convergence research that is driven by a specific and compelling societal problem and requires deep integration across disciplines and sectors. This training program provides researchers and practitioners with a foundational experience to position them for impact throughout their careers on the most challenging societal issues of our time. The curriculum gives participants the skillsets and networks to identify use-inspired problems and build convergence research solutions designed with intentional pathways to sustainable integration into existing societal systems.

Early 2023, in a program titled Tackling Climate-Induced Challenges with AI, the CORE Institute's training program, divided into two phases, consisted of a virtual six-week boot camp (conducted in spring 2023), where fellows engaged in a guided process of problem definition and solution ideation in cross-sector, cross-discipline teams. The curriculum comprised distinct segments for: (i) foundations of convergence research taught by selected instructors working on this area; (ii) case studies on AI strategies for climate-induced challenges; and (iii) hands-on breakout garage sessions for teams to engage in problem definition and new solution ideation.

In Phase II, supported by Institute mentors, a select group of participants attended a week-long, in-person summer incubator at UC San Diego. During this phase, fellows continued working together to draft proposals for funding and/or plans to prototype their AI solutions.

Local Students Use Advanced LEGOS in Robotics Workshop at San Diego Supercomputer Center



More than 50 local middle-school students used advanced LEGOS to engage in robotics-oriented science, technology, engineering and math (STEM) discovery over the summer at SDSC. San Diego Unified School District teacher Lori Holland, in partnership with SDSC through its StudentTech program, led the classes.

The science and robotics instructor at Marston Middle School explained that each day of the workshop increases the students' skill levels. "They first start out by building a simple robot out of LEGOs using instructions and by the end of day one or two, they have a solid understanding of basic coding through a variety of learning missions such as drive forward and turn in a circle, using sensors to make the robot 'smarter'. By day three or four, students are putting those basic skills to practice by using what they learned to complete a variety of missions in an actual FIRST LEGO League robot game," Holland said.

According to Holland, the students competed against each other by the end of the last day of camp to see who could score the most points in the robot game. High school mentors from around San Diego County assisted the students by providing programming guidance.

"The two weeks of classes revolved around LEGO EV3 Mindstorm robot sets, which came equipped with a central brick, entailing external connection ports, along with programmable sensors that allow the robot to react to external stimuli," said SDSC Education Manager Ange Mason. "Students utilized block coding software in order to instruct the robot to perform various tasks and our high school mentors helped Lori to ensure that each participant was able to clearly understand how this all worked."

This year's LEGO classes were just one of many summer programs offered by SDSC's StudentTech program. Among the 13 other workshops were programming bootcamps, career exploration seminars and a CAD class for students entering the fifth through twelfth grades. StudentTech has offered youth STEM programs for 17 years—including internships and summer camps.

"I think this program is a great way to introduce middle school students to robotics," Holland said. "This program is designed to get them excited and to seek out a FIRST LEGO League Robotics team in their neighborhood or school or to start their own competitive team."

Credit: Geisel Library by Erik Jepsen

The Dawning of a New School at UC San Diego

The final proposal for the School of Computing, Information, and Data Sciences (SCIDS) was voted on by the UC San Diego Academic Senate and passed with flying colors in October 2023. As an early next step, an assistant dean was appointed. Review by the University of California Board of Regents is expected during the summer of 2024. Once approved, the university will search for and appoint the new school's dean. The new institution will be founded around SDSC and the Halicioğlu Data Science Institute (HDSI). As the two foundational pillars for the new school, SDSC and HDSI will benefit from collaborations with other campus schools and academic departments, such as computer science and engineering, electrical and computer engineering, and cognitive science and mathematics. SDSC will serve as the operational and translational science core of the school, building on its history as one of the original four national supercomputer centers established by the U.S. National Science Foundation nearly 40 years ago. HDSI, which was established six years ago in anticipation of the growth of data sciences (SDSC Director Frank Würthwein and Chief Data Science Officer and CICORE Division Director Ilkay Altintas are founding faculty fellows), will serve as the academic core of the new school with an established undergraduate program, approved graduate degree programs and generous philanthropic support. The school will operate to meet the goal of transforming data into knowledge through development of data and information science, advancing innovative computing paradigms and developing new contextual learning algorithms and methodologies that can transform society. Its educational programs will be designed to train a new generation of professionals—particularly in the emerging technologies of artificial intelligence and machine learning. This all adds up to an exciting road ahead for the SDSC community. Stay tuned for future updates as a new school may be dawning.

San Diego Supercomputer Center
University of California San Diego
9836 Hopkins Drive
La Jolla, CA 92093-0505

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