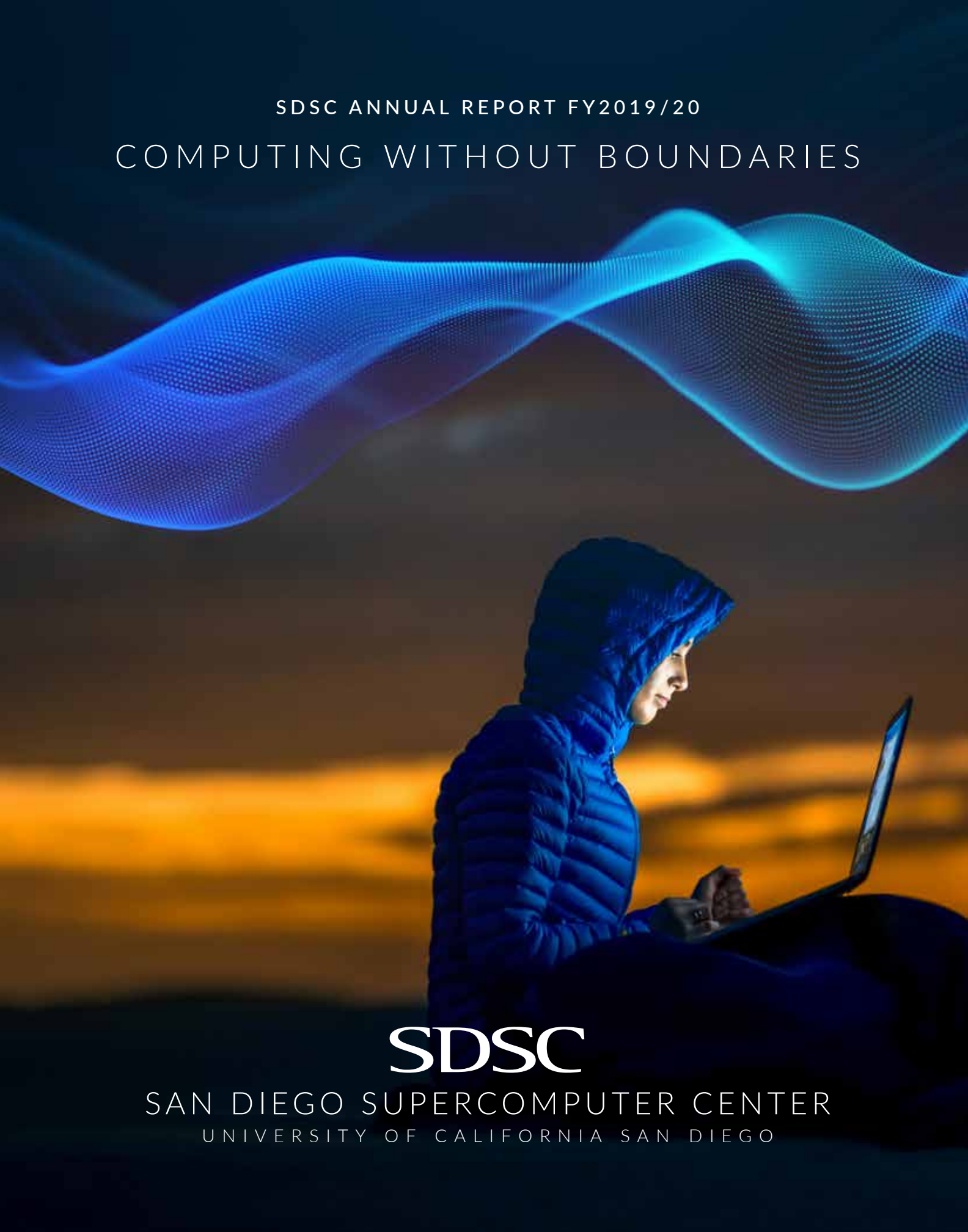
SDSC ANNUAL REPORT FY2019/20

COMPUTING WITHOUT BOUNDARIES



SDSC

SAN DIEGO SUPERCOMPUTER CENTER

UNIVERSITY OF CALIFORNIA SAN DIEGO

SAN DIEGO SUPERCOMPUTER CENTER
The San Diego Supercomputer Center (SDSC) is a leader and pioneer in high-performance and data-intensive computing, providing cyberinfrastructure resources, services, and expertise to the national research community, academia, and industry. Located on the UC San Diego campus, SDSC supports hundreds of multidisciplinary programs spanning a wide variety of domains from astrophysics and earth sciences to disease research and drug discovery. In late 2020 SDSC launched its newest National Science Foundation-funded supercomputer, *Expanse*. At over twice the performance of *Comet*, *Expanse* supports SDSC's theme of 'Computing without Boundaries' with a data-centric architecture, public cloud integration, and state-of-the art GPUs for incorporating experimental facilities and edge computing.

SDSC INFORMATION
Michael L. Norman, 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

Jan Zverina
Division Director, External Relations
jzverina@sdsc.edu
858-534-5111

COMPUTING WITHOUT BOUNDARIES
SDSC Annual Report FY2019/20
(PDF version available online at www.sdsc.edu/pub/)

EDITOR:
Jan Zverina

CO-EDITOR:
Kim Bruch

CONTRIBUTORS:
Kim Bruch, Julie Gallardo, Ron Hawkins, Fritz Leader, Susan Rathbun, Bob Sinkovits, Shawn Strande, Ben Tolo, Jan Zverina

CREATIVE DIRECTOR:
Ben Tolo

PHOTOGRAPHY:
Owen Stanley, Jon Chi Lou

All financial information is for the fiscal year that ended June 30, 2020. 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.

2 DIRECTOR'S LETTER
Adapting & Thriving During the New Normal

3 MEET MARK MILLER
"Pi Person" of the Year

ADVANCING COVID-19 RESEARCH
Page 4

IMPACT & INFLUENCE
Page 8

8 SDSC's National Mission in Advanced Cyberinfrastructure

16 State and UC Engagement

18 Campus and Education

SCIENCE HIGHLIGHTS
Page 23

24 **Human Health | Life Sciences**
From Combating Viruses to the Tree of Life

26 **Earth Sciences**
Understanding Fire & Ice through Data & Computation

28 **Exploring the Universe**
Supercomputer Simulations Reveal Secrets of the Universe

30 **Materials Engineering**
From Energy Savings to Improving our Environment

FOCUSED SOLUTIONS & APPLICATIONS
Page 33

34 Advanced Computation

40 Life Sciences Computing

42 Data-Driven Platforms and Applications

48 Data-Driven Disaster Relief

INDUSTRY RELATIONS
Page 52

BY THE NUMBERS: FY2019/20
Page 54

LEADERSHIP
Page 55

RESEARCH EXPERTS
Page 56

ADAPTING & THRIVING
DURING THE NEW NORMAL

To say that the period from mid-2019 to mid-2020, which is the bulk of what this Annual Report covers, was one of change is an understatement. The world shifted radically for all of us in March 2020 due to the COVID-19 pandemic, one year later, we're still grappling with the devastating effects of the coronavirus felt around the entire globe. Almost overnight, everyday habits and work patterns changed radically. The word 'zoom' took on a new meaning for most of us, as the world went virtual in so many respects.

But while so many things changed in our everyday lives, I soon realized that at SDSC, so many of us were already collaborating with other researchers remotely, so many of us were already adapting to different ways of communicating while maintaining our productivity and spirit of innovation. I want to again express my deep appreciation to those SDSCers and partners who kept our machine room running – in fact we even managed to stand up an all-new supercomputer, *Expanse*, which to my knowledge may be among the first HPC systems ever built during a global pandemic!

This year's report includes a section on COVID-19 research being conducted by SDSC staff as well as advancements enabled by SDSC resources and expertise. Researcher access to our National Science Foundation (NSF)-funded *Comet* supercomputer was coordinated through the COVID-19 HPC Consortium, a national alliance of computing resources announced in March 2020, and we expect *Expanse* to serve that research community after entering production in late 2020.

I'm pleased to report that SDSC remains in a strong position financially to weather the pandemic even as it lingers—possibly into 2021. Our success rate of winning federal and state grants and service contracts, plus industry funding, gives us a solid footing to plan ahead for the next few years. During the 2019-20 fiscal year SDSC had 77 active research grants awards totaling more than $71 million.

Another key NSF award in addition to *Expanse* is CloudBank, a five-year collaboration lead by SDSC to create a suite of managed services with the goal of simplifying public cloud access for computer science research and education. CloudBank entered production late last summer and I look forward to the benefits that this innovative public-private partnership will bring to the research and education community. Details are in the National Impact & Influence section starting on page 9.

Another new initiative for SDSC is the EarthCube Office (ECO). In late 2019 the NSF awarded SDSC and its partners a three-year, $5.9 million grant to host the EarthCube Office as part of the agency's EarthCube program, aimed at transforming geosciences research via an advanced cyberinfrastructure to further access and analysis of geosciences data and resources. This gives SDSC a leadership position in serving the needs of an entire NSF directorate's geosciences researchers, and like CloudBank, serves as a pathfinder for a new role for SDSC.

Last but not least, please join me congratulating Mark Miller for being named SDSC's 2020 'Pi' Person of the Year, a recognition that is long overdue given his outstanding work for many years. Early on, Mark explored the use of technology and robotics to enable high-throughput protein crystallization as part of the Joint Center for Structural Genomics. He has since focused on developing informatics tools to support biological/biomedical research, and created the CIPRES Science Gateway, a highly successful resource that provides browser-based access to high-performance computers for phylogenetics researchers. Mark and his colleagues were recently awarded a $1 million NSF grant to create the X-ray Imaging of Microstructures Gateway, or XIMG, so that material sciences researchers can study new and existing materials using X-ray diffraction. Read more about Mark's accomplishments on the next page.

I invite you to browse through our latest Annual Report, which shows the depth and breadth of all we do at SDSC to help advance scientific discovery. Please stay healthy and safe, and may the upcoming year bring a better world for all of us!

Michael L. Norman
SDSC Director

MEET MARK MILLER: SDSC'S 2020 'PI' PERSON OF THE YEAR

SDSC Biologist Mark Miller was named the Center's 2020 π Person of the Year. Now in its seventh year, the award recognizes SDSC researchers who have one leg, so to speak, in a science domain and the other in cyberinfrastructure technology.

"Mark has worked tirelessly to make the CIPRES Science Gateway the most impactful one running on XSEDE resources today," said SDSC Distinguished Scientist Wayne Pfeiffer, a colleague of Miller's. "To date, tens of thousands of users from around the world have used CIPRES, and recently, more than 100 users from nearly 30 countries have used it to study the COVID-19 pandemic."

Miller joined SDSC in 2000 to explore the use of technology and robotics to enable high-throughput protein crystallization as part of the Joint Center for Structural Genomics. He founded the Next Generation Tools for Biology Group at SDSC in 2003 and has worked since then to develop informatics tools and software that support biological/biomedical research. As part of that effort, he created the CIPRES Science Gateway, a highly successful online resource that provides browser-based access to high-performance computers for phylogenetics researchers.

With expertise in molecular and structural biology, biochemistry, and bioinformatics, Miller has sustained CIPRES for the past decade as a principal investigator (PI) on grants from both the National Science Foundation and the National Institutes of Health (NIH). During the past year he received an award from Internet2 to enable CIPRES to submit to commercial clouds. The project showed how cloud computing can improve both capacity and turnaround time for CIPRES jobs.

Miller received his B.S. in Biology from Eckerd College in 1976 and his Ph.D. in Biochemistry from Purdue University in 1984. He came to UC San Diego for a postdoctoral position, and subsequently became a research scientist at Monsanto/Kelco in San Diego. Through his work in these positions, he developed expertise in many areas of biology, publishing more than 40 papers in areas including phospholipid and steroid metabolism, eukaryotic cell biology, protein crystallography, electron transfer kinetics, metabolic engineering, and more.

"Working at SDSC gave me access to experts in high-performance computing such as Wayne Pfeiffer, and placed me with a powerful team of advocates for Science Gateway creation, including Nancy Wilkins-Diehr and Michael Zentner," said Miller. "This made it possible for me as a domain biologist to create a production-quality cyberinfrastructure resource that allows thousands of researchers each year to easily analyze data on large compute clusters."

In 2020 Miller and a multidisciplinary team of researchers at the University of Minnesota, Carnegie Mellon University, and Cornell University were awarded a $1 million RAISE (Research Advanced by Interdisciplinary Science and Engineering) grant from the NSF to create the X-ray Imaging of Microstructures Gateway (XIMG), a science gateway designed to make it possible for global materials science researchers to study the behavior of new and existing materials using high- and low-field X-ray diffraction.

"The XIMG (pronounced X-image) will be the first of its kind for the materials science community where toolkits are available for visualization, modeling, and simulation at mesoscale and nanoscale levels," said Miller. *(Read more about CIPRES on page 13)*

Watch how researchers use the CIPRES Science Gteway in the YouTube video "Shining Light on Landfills to Uncover 'Dark Life'" using the above QR code or url.

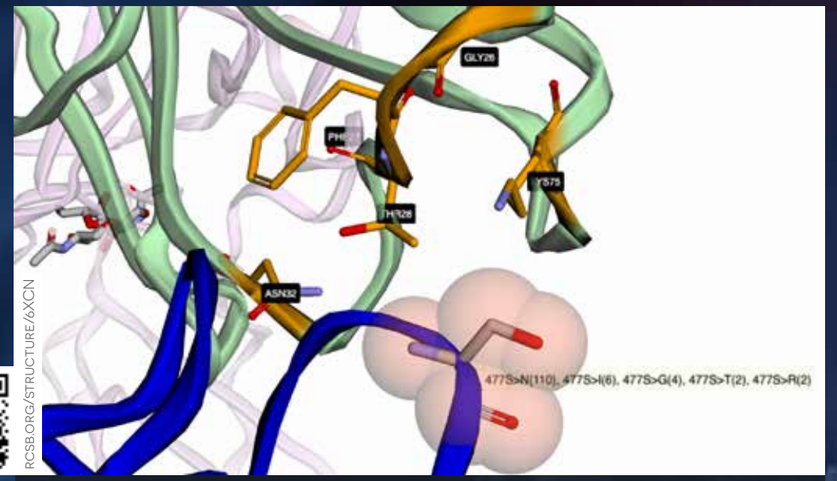
COVID19 RESEARCH

PROVIDING **RESOURCES** & **EXPERTISE** TO ADVANCE **COVID-19 RESEARCH**

As the U.S. and other countries worked 24/7 to mitigate the devastating effects of the COVID-19 disease caused by the SARS-CoV-2 virus, SDSC contributed by providing researchers priority access to its high-performance computing (HPC) systems and other resources to advance our understanding of the virus. Access to the Center's National Science Foundation (NSF)-funded *Comet* supercomputer was coordinated through the COVID-19 HPC Consortium, a national alliance of computing resources announced in March 2020. The consortium combines the capabilities of some of the most powerful computers in the world – 600+ petaflops of compute power plus extensive cloud resources.

"Supercomputers have already demonstrated their capabilities in accelerating scientific research, and like no other event before, this pandemic has underscored their importance in benefiting science and society," said SDSC Director Michael Norman. "For us, it crystalized SDSC's mission, which is to deliver lasting impact across the greater scientific community by creating innovative end-to-end computational and data solutions to meet the biggest research challenges of our time."

Numerous researchers at SDSC, UC San Diego, and other academic institutions are also involved in several COVID-19 research projects at SDSC. Some highlights follow.



RCSB.ORG/STRUCTURE/6XCN

Observed mutations at position 477 (pink spheres) in the SARS-CoV-2 spike glycoprotein (blue) mapped to the binding interface (orange) of the C105 neutralizing antibody Fab fragment (green) in PDB structure 6XCN. *Credit: Peter Rose, SDSC/UC San Diego.*

FIGHTING COVID-19 WITH KNOWLEDGE GRAPHS

In May 2020 the NSF awarded SDSC researchers funding to organize COVID-19 information into a transdisciplinary knowledge network that integrates health, pathogen, and environmental data to better track cases to improve analysis and forecasting across the greater San Diego region. The award calls for a quick launch of a comprehensive semantic integration platform for data-driven analysis and development of policy interventions, taking into account social, economic, and demographic characteristics of populations in different areas and biomedical data, such as virus strains and genetic profiles, according to Peter Rose, director of SDSC's Structural Bioinformatics Laboratory and principal investigator (PI) for the project. Ilya Zaslavsky, director of Spatial Information Systems Laboratory at SDSC and UC San Diego, is a co-PI of the award. The project team has planned for continued collaboration with industry, university, and community efforts focused on COVID-19 analysis, including government agencies in San Diego County, citizen scientists (Open San Diego), and industrial partners such as Microsoft and Neo4J.

The project incorporates viral genome data for 70,000+ SARS-CoV-2 strains via the COVID-19-Net Knowledge Graph, and hosts a set of Jupyter Notebooks with built-in 3D visualizations to analyze the effect of mutations on the binding of small molecules, nanobodies, and antibodies. Researchers can adopt and extend these Notebooks for their own research on the development small molecule therapeutics and vaccines for COVID-19.

Knowledge graphs were also constructed using SDSC's AWESOME (Analytical Workbench for Exploring SOcial Media) platform, which harnesses the latest 'big data' technologies to collect, analyze, and understand social media activity along with current events data and domain knowledge. The National Institutes of Health (NIH)-funded TemPredict project seeks to develop machine learning models to predict the onset of COVID-19 in an individual by integrating multiple physiological data from wearable sensors with health surveys. *(Read more about AWESOME on page 44)*

QRGO.PAGE.LINK/K4EUJ

QRGO.PAGE.LINK/YU9SM

COVID19 RESEARCH
Using SDSC's Resources & Expertise

TEMPREDICT COVID-19 EARLY DETECTION SYSTEM

SDSC participates in the TemPredict collaborative between UC San Diego and UC San Francisco, funded by the Department of Defense and wearable device providers. The initiative is led by Benjamin Smarr from UCSD Bioengineering and the Halicioğlu Data Science Institute, and SDSC Chief Data Science Officer Ilkay Altintas.

On March 20, 2020, the team launched TemPredict, which during 2020 grew to a 60,000+ participant study developing algorithms to detect COVID-19 infection. The initiative collects continuous physiology data from wearable devices, and trains machine learning on these data using descriptive and daily surveys covering symptoms and diagnoses. TemPredict's sample population includes participants who possessed an Oura ring prior to joining the study as well as ~4,000 healthcare workers to whom TemPredict provided the rings.

"The COVID-19 pandemic has highlighted the urgent need to develop solutions that identify individuals who may be in the earliest stages of an illness, and to provide actionable intelligence to those individuals as well as those orchestrating larger responses," said Smarr.

The algorithms are being developed to detect onset at or before the time of first symptom report. The goal is to provide early alerts to those likely to be infected so that they can seek testing and respond accordingly. "TemPredict supports a better understanding of COVID-19 by focusing on the data from individuals and combining it with population level studies," said Altintas, who also leads WIFIRE, which serves as a foundational system for TemPredict. The study also makes use of SDSC's AWESOME platform and its Sherlock secure cloud project.

QRGO.PAGE.LINK/HUPZ6

USING MOLECULAR DYNAMICS AND MACHINE LEARNING FOR VIRUS MODELING AND SCREENING

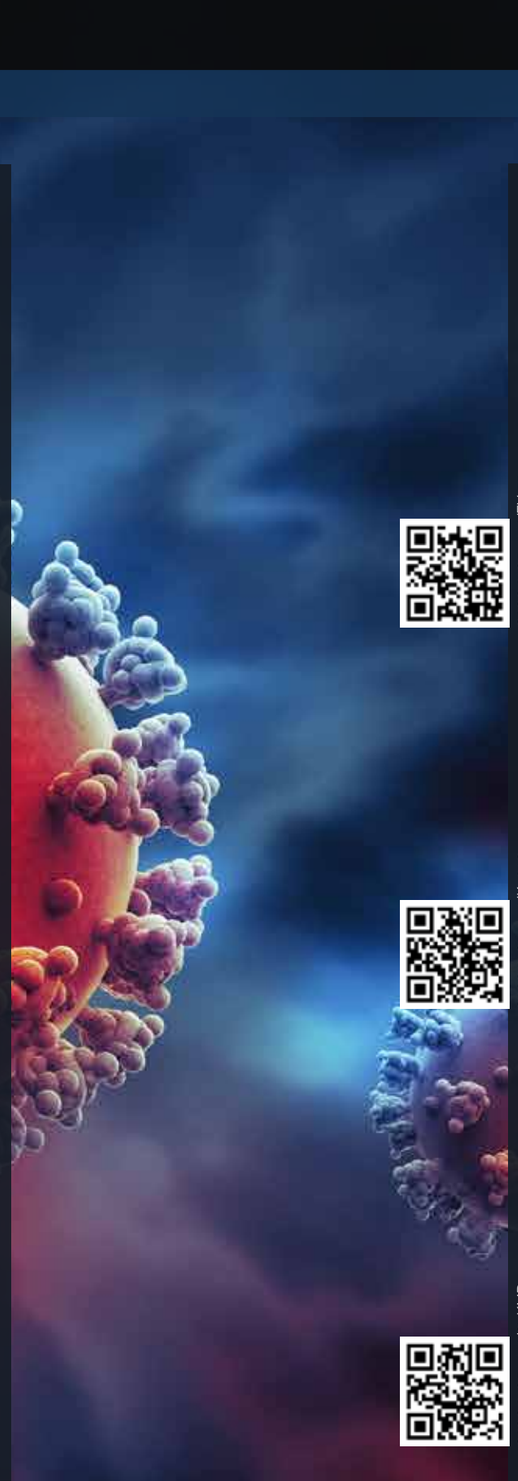
Accurate structures of protein targets in SARS-CoV-2 are needed as the starting point for computational screening efforts toward identifying effective treatments of COVID-19 through drug repurposing or novel drug design. Michael Feig, a professor of biochemistry and molecular biology and chemistry at Michigan State University, and his postdoctoral colleague Lim Heo, focused on the four membrane proteins nsp4, nsp6, M, and E and began with initial models resulting from machine-learning based methods that were generated by them. The initial models were subsequently refined via extensive molecular dynamics (MD) simulations with the goal of generating high-accuracy computational models for more virus model building and screening. The MD-based refinement method involved extended simulations using the GPU (graphic processing unit) capabilities of SDSC's *Comet* supercomputer.

QRGO.PAGE.LINK/AMK7G

PREDICTIVE SCIENCE INC. USES COMET TO STUDY COVID-19 DATA PATTERNS

A researcher with San Diego-based Predictive Science Inc. has been using SDSC's *Comet* supercomputer to study COVID-19 data patterns using modified influenza data patterns. "We have been basically combining our models of influenza and our models of COVID, so we can run the calculations and train the models and test them to see which ones do a better job at forecasting," said Michal Ben-Nun, who has a background in infectious diseases. "Once we have a working code that is fully debugged, we will have to start running that on either *Expanse* or on *Comet*."

YOUTU.BE/QMM3HIPCHVU



GO FAIR U.S. COORDINATION OFFICE ASSISTS IN VIRUS DATA COLLECTION

GO (Global Open) FAIR is a 'bottom up' initiative aimed at implementing the FAIR principles to ensure that data is findable, accessible, interoperable, and reusable. SDSC's Research Data Services division hosts the U.S. GO FAIR coordination office. To improve global virus data collection and sharing, in 2020 GO FAIR implemented the Virus Outbreak Data Network (VODAN), which is being integrated with other data sources. "Our short-term goal was to create ways for collecting and sharing COVID-19 information, which can then be integrated with other data sources for visibility of the virus across broad regions and borders," said Christine Kirkpatrick, RDS division director. "For example, it's possible for a rural clinic to establish a data point that shares limited summary information about patients being treated, and more detail with trusted partners such as governmental health organizations. This allows a country and/or county-wide data stewards and scientists to analyze data points and more quickly draw inferences to inform policymakers and health officials." In partnership with VODAN Africa and the West Hub, GO FAIR US moderated a three-part webinar series on the role of data stewardship to better understand the spread of COVID-19 in partner countries.

QRGO.PAGE.LINK/FYJFF

ACCELERATING COVID-19 RESEARCH BY OPTIMIZING GPU PERFORMANCE

SDSC researchers applied their high-performance computing expertise by porting the popular UniFrac microbiome tool used in the Human Microbiome Project to GPUs to increase the acceleration and accuracy of scientific discovery, including COVID-19 research. Microbiomes are the combined genetic material of the microorganisms in a particular environment, including the human body. Igor Sfiligoi , SDSC's lead scientific software developer for high-throughput computing, collaborated with Rob Knight, founding director of the Center for Microbiome Innovation and a professor of Pediatrics, Bioengineering and Computer Science & Engineering at UC San Diego; and Daniel McDonald, scientific director of the American Gut Project. "This work did not initially begin as part of the COVID-19 response," said Sfiligoi. "We started the discussion about such a speed-up well before, but UniFrac is an essential part of the COVID-19 research pipeline."

QRGO.PAGE.LINK/KN24C

DETECTING POTENTIAL COVID-19 PROTEASE INHIBITORS

SDSC-affiliated researchers Valentina Kouznetsova and Igor Tsigelny recently created a pharmacophore model and data mined the conformational database of FDA-approved drugs that identifies 64 compounds as potential inhibitors of the COVID-19 protease. Among the selected compounds were two HIV protease inhibitors, two hepatitis C protease inhibitors, and three drugs that have shown positive results in testing with COVID-19.

The conformations of these compounds underwent three-dimensional fingerprint similarity clusterization. The researchers also conducted docking of possible conformers of these drugs to the binding pocket of protease and then conducted the same docking of random compounds.

"We published our findings with Chemrxiv so we could move on with proper testing that would allow this concept to be used during the pandemic crisis," said Tsigelny, adding that the study's key point is elucidation of FDA-approved drugs that can be quickly tried and used on patients, an avenue supported by other COVID-19 research.

QRGO.PAGE.LINK/GXKMD

IMPACT & INFLUENCE

SDSC'S **NATIONAL MISSION** IN **ADVANCED CYBERINFRASTRUCTURE**

As one of the first four U.S. supercomputer centers opened in 1985 by the NSF, SDSC has an impressive history of programs and partnerships that have benefited science and society across an ever-increasing number of research domains. SDSC's mission has expanded in recent years to encompass more than just advanced computation, which has served as a foundation to include innovative applications and expertise related to the growing amount of digitally-based science data generated by researchers. *(See page 35 for more on SDSC's new Expanse supercomputer)*

SDSC Shares Top Supercomputing Achievement Award

ORGO.PAGE.LINK/UTQ4D

In late 2019 SDSC received three top HPCwire honors, including the online publication's Readers' Choice Award for the use of the Center's *Comet* supercomputer in helping astrophysics researchers gain new insights into gravitational waves, or invisible space ripples, via supercomputer simulations. HPCwire recognized SDSC along with researchers at the Perimeter Institute for Theoretical Physics in Ontario, Canada; and the Theoretical Astrophysics Program at the University of Arizona. Also supporting this landmark research project was the NSF's Extreme Science and Engineering Discovery Environment (XSEDE) program, which allocated time on HPC systems at other supercomputing centers. *(See pages 32 and 40 and for additional HPC awards)*

CloudBurst Team Receives CENIC Innovations Award

ORGO.PAGE.LINK/CBWFV

The California Research and Education Network (CENIC) and the Pacific Research Platform (PRP) awarded researchers at SDSC and the Wisconsin IceCube Particle Astrophysics Center (WIPAC) at the University of Wisconsin – Madison its 2020 'Innovations in Networking Award for Experimental Application' for their bold experiment last November that marshalled all globally available-for-sale GPUs across Amazon Web Services, Microsoft Azure, and the Google Cloud Platform.

ORGO.PAGE.LINK/UTQ4D

In November 2019, about 51,500 GPU processors were used in the experiment, which demonstrated that IceCube can effectively use a large number of GPUs in a single pool. The research team included Frank Würthwein, SDSC lead for high-throughput computing (HTC); Igor Sfiligoi, SDSC's lead scientific software developer for HTC; Benedikt Riedel, global computing coordinator for the IceCube Neutrino Observatory and computing manager at WIPAC; and David Schultz, a production software manager with IceCube. "We showed that a cloud-based cluster can provide almost 90 percent of the performance of the *Summit* supercomputer at Oak Ridge National Laboratory, at least for the purpose of IceCube simulations," said Sfiligoi.

Shava Smallen is co-principal investigator and lead architect for the CloudBank user portal.

CloudBank Now Operational

ORGO.PAGE.LINK/PQMW9

In September 2020, SDSC and its partners at the University of Washington, UC Berkeley, and Strategic Blue entered production operations of the NSF-funded CloudBank program, which aims to simplify the use of public clouds across computer science research and education. CloudBank is the result of a mid-2019 $5 million NSF award to SDSC and its partners to develop a suite of managed services to simplify public cloud access for computer science research and education.

The CloudBank team has been working since the summer of 2019 with public cloud providers, early users, project advisors, and the broader computer science community to develop the processes, tools, and education and outreach materials that address the plethora of pain points in making effective use of public clouds. These initiatives included developing a broad understanding of capabilities available from providers, determining cost estimates for inclusion in NSF proposals, and assisting researchers in accessing the cloud and carrying the full range of user management and usage tracking that is critical for successful research and education programs.

"This transition to production operations is an important milestone for the project and represents the culmination of a lot of hard work by all our partners this past year," said Michael Norman, SDSC director and CloudBank's principal investigator. "I look forward to the benefits that this innovative public-private partnership will bring to the research and education community."

"We designed the CloudBank portal so that researchers would have a simplified interface to compare and access a variety of cloud provider services and monitor their spending across them in a unified dashboard," said Shava Smallen, co-principal investigator and lead architect for the CloudBank user portal.

NATIONAL IMPACT&INFLUENCE

Highlights of Key National Partnerships

EXTREME SCIENCE AND ENGINEERING DISCOVERY ENVIRONMENT (XSEDE)

XSEDE.ORG

The NSF's XSEDE program allows scientists to interactively access and share computing resources, data, and expertise. SDSC's *Comet* and new *Expanse* supercomputers are accessible via the XSEDE allocation process to U.S. researchers as well as those affiliated with U.S.-based research institutions.

OPEN SCIENCE GRID CONSORTIUM

ORGO.PAGE.LINK/KNBJO

OSG is a multidisciplinary research partnership dedicated to the advancement of all open science via the practice of distributed High Throughput Computing (dHTC). The NSF recently funded the 'Partnership to Advance Throughput Computing (PATh)', a collaboration between the OSG Consortium and the Center for High Throughput Computing (CHTC) at the University of Wisconsin - Madison. PATh is a 5-year project funded by the NSF's Office of Advanced Cyberinfrastructure's Campus Cyberinfrastructure (CC*) program to address the needs of the rapidly growing community of faculty and students who are embracing dHTC technologies and services to advance their research. Through a partnership with XSEDE, OSG scientists use PATh services to access resources such as *Expanse* and *Comet* to further their research.

SUPPORTING THE NATIONAL BRAIN INITIATIVE THROUGH THE NEUROSCIENCE GATEWAY

WWW.NSGPORTAL.ORG

Charting brain functions in unprecedented detail could lead to understanding how the brain functions and new prevention strategies and therapies for disorders such as Alzheimer's disease, schizophrenia, autism, epilepsy, traumatic brain injury, and more. The BRAIN Initiative (Brain Research through Advancing Innovative Neurotechnologies), launched by President Barack Obama in 2013, is intended to advance the tools and technologies needed to map and decipher brain activity, including advanced computational resources and expertise.

EARTHCUBE OFFICE

EARTHCUBE.ORG

The EarthCube Office (ECO) spent much of 2019 and 2020 expanding EarthCube's reach across an array of geosciences disciplines, while helping to coordinate several new cyberinfrastructure systems that support EarthCube's goal of creating a well-connected environment to share data and knowledge in an open and inclusive manner, accelerating our ability to better understand and predict the Earth's systems. ECO supports EarthCube community-led governance and provides technical and coordination support for EarthCube funded projects. Strabospot is only one example of a useful EarthCube tool revitalized in 2020. Originally created for field geologists to collect and record field data, StraboSpot was significantly modified to assist education communities to meet their needs due to pandemic restrictions.

SHERLOCK

ORGO.PAGE.LINK/GRB8P

In mid-2019 SDSC's Sherlock division launched its Innovation Accelerator Platforms within its Sherlock Cloud infrastructure and its newest offering, Vyloc Cloud, a managed cloud capability for open (non-protected) data. In June 2020 the division expanded its multi-cloud solution, Sherlock Cloud, to include the Google Cloud Platform (GCP). The expansion secures the trifecta of major public commercial cloud platforms included within Sherlock's secure compute and data management solution offerings. *(Read more about Sherlock on page 46)*

OPEN STORAGE NETWORK (OSN)

OPENSTORAGENETWORK.ORG

As wildfires blazed throughout California in mid-2020, SDSC's Research Data Services (RDS) team used the OSN to assist several UC Santa Cruz entities with urgent needs for offsite backup as power outages threatened critical data loss across campus. As an NSF-funded distributed data storage demonstration project, OSN allowed the team to quickly assist with transferring active research data. In late 2020 and Spring 2021, the OSN hosted a four-part webinar series on the emerging needs related to research data.

WEST BIG DATA INNOVATION HUB (WBDIH)

WESTBIGDATAHUB.ORG

The NSF-funded West Big Data Innovation Hub's mission is to build and strengthen partnerships across industry, academia, nonprofits, and government to address societal and scientific challenges, spur economic development, and foster a national 'big data' ecosystem. Recent activities included a four-part webinar series on COVID-19 and data science. In partnership with the Border Solutions Alliance (BSA), the Hub held a COVID-19 data challenge that asked teams to look at data-driven ways for describing and improving life along the border. Teams included BSA partners from both sides of the US-Mexico border, and included a category for high school students. The Hub serves the 13 western states from Montana to New Mexico and everything west, including Hawaii and Alaska. Thematic areas include metro/urban data science as well as natural resources and hazards, with a focus on water. Cross-cutting areas include cloud computing, data challenges, and storytelling communities of practice, public policy and ethics, and responsible data science including sharing and security. SDSC Director Michael Norman is the WBDIH's principal investigator, with Christine Kirkpatrick, who leads SDSC's Research Data Services division, as co-PI.

GO FAIR US COORDINATION OFFICE

ORGO.PAGE.LINK/MJRYA

SDSC's Research Data Services division continues to host the U.S. GO FAIR coordination office. GO (Global Open) FAIR is a 'bottom up' initiative aimed at implementing the FAIR principles to ensure that data is findable, accessible, interoperable, and reusable. In 2020, in an effort to improve virus data collection and sharing around the world, GO FAIR implemented the Virus Outbreak Data Network (VODAN), which is being integrated with other data sources. VODAN's immediate goal is to create ways for collecting and sharing COVID-19 information in place, which can then be integrated with other data sources for visibility of the virus across broad regions and borders. In partnership with VODAN Africa and the West Hub, GO FAIR US moderated a three-part webinar series in May on the role of data stewardship to understand the spread of COVID-19 in partner countries.

UNIVERSAL SCALE STORAGE (USS)

ORGO.PAGE.LINK/SC7ML

In 2019, SDSC's Research Data Services division launched the Universal Scale Storage (USS) service for UC San Diego, the UC system, and the local research community. The service hosts several petabytes of data, and provides a single scalable namespace directly mounted on campus research lab systems, SDSC supercomputers, and compute clusters hosted in SDSC's data center. Recognizing extensive research demand by a global audience to interact with USS, RDS created the USS Scale Storage Cloud Connector, which extends local access methods outside the UC San Diego campus and enables cloud-native applications with USS.

AMERICAN ASSOCIATION FOR THORACIC SURGERY ADOPTS HUBZERO® CLOUD PLATFORM

In early 2020 the American Association for Thoracic Surgery (AATS) adopted an open-source, cloud-based platform led out of SDSC that addresses widely recognized challenges with historical platforms throughout the cardiothoracic surgical community. The new AATS Quality Assessment Platform is built on HUBzero, an open-source software platform that hosts analytical tools, publishes data, shares resources, and builds collaborative communities in a single web-based ecosystem. The AATS platform is a single science gateway that includes customizable dashboards and advanced data visualization while providing hospitals and physicians with unfettered access to their data and the ability to share best practices and establish new research collaborations. The AATS platform also includes risk adjustment models using machine learning techniques developed in collaboration with SDSC. "The AATS Quality Assessment Platform is a real-time, collaborative analytics platform for all thoracic specialties, with the goal of delivering significant improvements across the board, notably improved patient care," said Michael Zentner, the director of HUBzero who joined SDSC as director of sustainable scientific software in June 2019 following nine years with Purdue University, where he was an entrepreneur-in-residence at the Purdue Foundry and a senior research scientist.

ORGO.PAGE.LINK/WMBUV

PACIFIC RESEARCH PLATFORM GOES GLOBAL

ORGO.PAGE.LINK/BPYXY

To meet the needs of researchers in California and beyond, in 2015 the NSF awarded a five-year grant, now extended to a sixth year, to fund the Pacific Research Platform (PRP), a high-speed data transfer network led by researchers at UC San Diego and UC Berkeley that now connects more than 50 institutions including the Department of Energy and multiple research universities around the nation and the world. The PRP has since evolved into what is today a nascent national research platform, and is now also part of the Global Research Platform (GRP) under the guidance of Principal Investigator Larry Smarr, a UC San Diego computer science and engineering professor as well as founder of the California Institute for Telecommunications and Information Technology (CALIT2) at UC San Diego. While Smarr retired in June 2020, SDSC staff involved with the PRP project include Frank Würthwein, a PRP co-PI and lead for distributed high-throughput computing; and Thomas Hutton, a networking architect. Phil Papadopoulos, formerly director of cloud and cluster software development at SDSC, is also a PRP co-PI.

ORGO.PAGE.LINK/3YTBX

Michael Zentner is director of HUBzero and principal investigator for SGCI.

SCIENCEGATEWAYS.ORG

PROVIDING 'SCIENCE GATEWAYS' FOR RESEARCHERS

In mid-2016, a collaborative team led by SDSC Associate Director Nancy Wilkins-Diehr was awarded a five-year, $15 million NSF grant to establish a Science Gateways Community Institute (SGCI) to accelerate the development and application of highly functional, sustainable science gateways that address the needs of researchers and educators across the full spectrum of federally funded programs. Upon Nancy's retirement, Michael Zentner assumed the project lead in mid-2019. Science gateways make it possible to run scientific applications on supercomputers such as *Expanse* with easy-to-use browser access that allows researchers to focus on their scientific endeavors without having to learn the details of how supercomputers work. Science gateways also establish learning environments for students, share large data sets, create collaborations, and disseminate large volumes of research products.

Now completing its fourth year, the SGCI has served over 100 clients with consultations, produced 285 publicly available research products, and had more than 2,000 attendees at its events. SGCI has a strong commitment to diversity, with 382 out of 511 faculty and students serving in its workforce development efforts belonging to underrepresented populations. The scientific impact of the clients served is evident: SGCI client gateways have been cited more than 41,000 times in the literature. The value proposition of SGCI services is strong, with exit interviews from SGCI clients documenting a speedup factor of 7.6 times over the amount of effort clients would need to expend to achieve similar results. The community recognizes this, with members having written into their proposals more than $4 million in funding this year for SGCI personnel to help with their gateway efforts. SGCI's next steps started in 2020 are Tech Summit activities targeted at enabling interoperability between gateways and the pursuit of long-term sustainability plans.

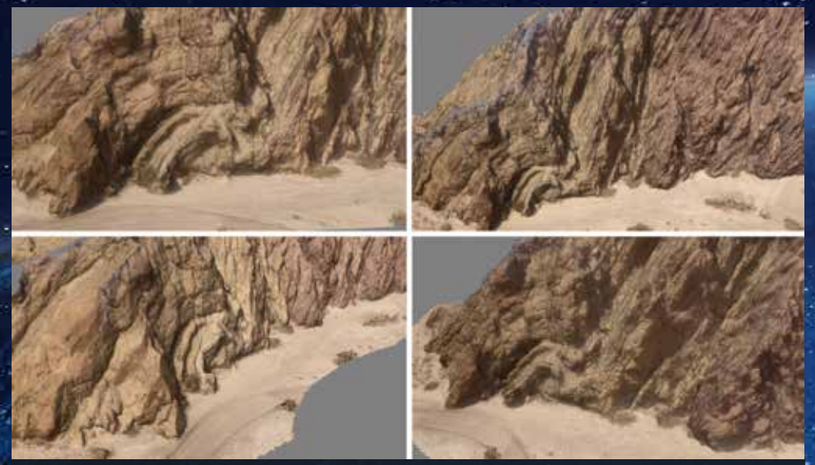
Mark Miller is principal investigator for the CIPRES science gateway.

PHYLO.ORG

SCIENCE GATEWAYS PIONEERED BY SDSC RESEARCHERS

CIPRES

One of the most popular science gateways across the entire XSEDE resource portfolio is the CIPRES science gateway, created as a portal under the NSF-funded CyberInfrastructure for Phylogenetic RESearch (CIPRES) project in late 2009. Based at SDSC, CIPRES supports more than 10,000 researchers each year who are investigating a wide range of topics about life on earth, including how viruses and bacteria mutate through time, how many species are there are and how they work together to create functional ecosystems, and how animals and plants evolve and adapt to a changing environment. Currently, CIPRES runs parallel versions of 21 widely used sequence alignment, tree inference, and population biology codes on SDSC's *Comet* supercomputer, which has both CPU cores and GPUs. In 2019 CIPRES was awarded a one-year Internet2 grant funded by NSF that allowed gateway users to take advantage of more powerful compute processors available from a commercial cloud provider AWS to accelerate their scientific discoveries. By mid-2020, CIPRES has supported more than 35,000 scientists in producing 7,500 peer-reviewed papers published in journals such as *Nature*, *Cell*, *Science*, and *Proceedings of the National Academy of Sciences*, and delivered more than 140 million equivalent-core hours of compute time to scientists. *(Read more about Mark Miller's new award for a materials science gateway on Page 3)*



To address the need for continuing educational resources for students during the pandemic, OpenTopography created a virtual field activity where students learn many of the same skills in field camp virtually. In this exercise, students analyze a colored point cloud of geologic folded or bent rocks at Painted Canyon adjacent to the San Andreas Fault in Mecca Hills near Palm Springs, California. Similar to how they would in a field camp, the students virtually collect, plot, and analyze their measurements of the rocks to unravel the history of a major tectonic fault in California. *Credit: Chelsea Scott, Arizona State University.*

OPENTOPOGRAPHY COLLABORATION AWARDED NEW FOUR-YEAR GRANT

In early 2020 NSF renewed funding for OpenTopography, a science gateway that provides online access to Earth science oriented high-resolution topography data and processing tools to a broad user community advancing research and education in areas ranging from earthquake geology to ecology and hydrology. The award, jointly funded by the Geoinformatics and the Geomorphology and Land Use Dynamics programs in the Division of Earth Sciences at NSF, provides $2.55 million over four years for the fourth generation of the project. Founded in 2009, OpenTopography is managed by SDSC in collaboration with UNAVCO and Arizona State University's School of Earth and Space Exploration.

Viswanath Nandigam is principal investigator and chief software architect for the OpenTopography Facility.

The new award allows OpenTopography to accelerate data-intensive topography-based research while addressing scalability, adding new features and functionality, and expanding data offerings. This project also enhance access to data and processing tools that improve efficiency and economic competitiveness. "OpenTopography has matured into a widely used platform that makes topographic data easy to discover, access, and use," said Christopher Crosby, an OpenTopography principal investigator and Geodetic Imaging project manager at UNAVCO.

"OpenTopography continues to experience very strong growth in the number of users and jobs being run," said Viswanath Nandigam, the project's principal investigator and associate director for the Advanced Cyberinfrastructure Development group at SDSC. "A major effort during this phase is a migration to the commercial cloud with its auto scaling, elastic load balancing, and resilient architectures to better handle the needs of the growing community."

ORGO.PAGE.LINK/W5MGE

NATIONAL IMPACT&INFLUENCE
SCIENCE GATEWAYS PIONEERED BY SDSC RESEARCHERS

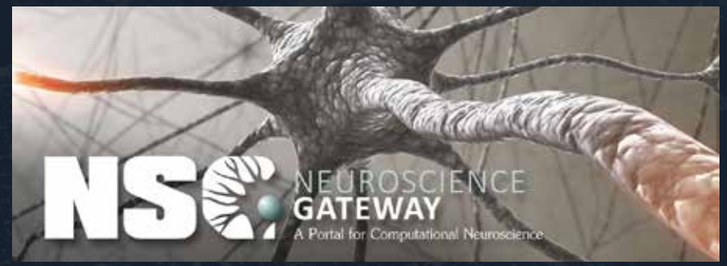
NANOHUB.ORG

HUBZERO.ORG

SDSC HEADQUARTERS HUBZERO® SCIENCE GATEWAY PLATFORM

During 2019, the HUBzero science gateway platform transitioned to its new headquarters provided by SDSC. The platform has been used to construct science gateways starting in 2007 with nanoHUB.org, and today operates more than 20 gateways from SDSC data centers provided by RDS and Sherlock. HUBzero is used in many scientific domains ranging from nanotechnology to human-animal bonding, and serves audiences spanning high school students through advanced university researchers. The platform operates at scale, providing service to two million unique individuals annually. The capabilities of HUBzero include middleware for connecting to high-performance computing (HPC) and high-throughput computing (HTC) systems, a simulation tool hosting environment that allows users to interact with powerful simulations using only a web browser, collaboration tools, and a publishing pipeline. HUBzero is also an open source platform being used by many other gateways outside of those operating at SDSC. The platform is led by Michael Zentner, SDSC's director of sustainable scientific software. With its unique revenue model, HUBzero marks the first product from the Center's sustainable scientific software initiative.

Amit Majumdar is principal investigator for the Neuroscience Gateway.



NEUROSCIENCE GATEWAY

The Neuroscience Gateway (NSG) facilitates access and use of NSF-funded HPC resources by neuroscientists. Amit Majumdar, director of SDSC's Data Enabled Scientific Computing (DESC) division, is NSG's principal investigator, with Subhashini Sivagnanam, a principal scientific computing specialist with DESC as one of the project's co-PIs. The NSG is also exploring incorporation of HTC and academic and commercial cloud computing resources. Computational modeling of cells and networks has become an essential part of neuroscience research, and investigators are using models to address problems of ever-increasing complexity, such as large-scale network models and optimization or exploration of high dimensional parameter spaces. In addition, in recent years neuroscientists are utilizing NSG to process EEG, MRI, and fMRI data using the data processing software provided by NSG, as well as using the machine learning tools provided by NSG. NSG catalyzes such research by lowering or eliminating the administrative and technical barriers that currently make it difficult for investigators to use HPC resources. It offers free computer time to neuroscientists acquired via the supercomputer time allocation process managed by the Extreme Science and Engineering Discovery Environment (XSEDE) Resource Allocation Committee (XRAC). NSG provides access to popular neuroscience tools, pipelines, data processing software installed on various computing resource while providing a community mailing list for neuroscientists to collaborate and share ideas.

NSGPORTAL.ORG

IMPACT & INFLUENCE

STATE AND UC ENGAGEMENT

FORGING RESEARCH PARTNERSHIPS

UC CAMPUSES COLLABORATE TO COMBAT COVID-19

In May 2020 the National Science Foundation (NSF) awarded two SDSC researchers funding to organize COVID-19 information into a transdisciplinary knowledge network that integrates health, pathogen, and environmental data to better track cases to improve analysis and forecasting across the greater San Diego region. The COVID-19-Net grant was one of 13 RAPID awards made by the NSF Convergence Accelerator to Track A: Open Knowledge Network-related projects. This project is being coordinated with another RAPID program led by Krzysztof Janowicz, a professor of Geographic Information Science at UC Santa Barbara, with focuses on infrastructure resilience, supply chain disruptions, and local policy decisions designed to combat the pandemic. UC San Francisco also is a partner in the Open Knowledge Network. *(Read more about SDSC-enabled COVID-19 research on page 4, and UC Berkeley's involvement in CloudBank on Page 9)*

HPC@UC

This novel program provides UC researchers access to SDSC's supercomputing resources and expertise. Since its 2016 inception HPC@UC has assisted some 50 individual research projects spanning all ten UC campuses, with over 20 million core-hours of compute time allocated on SDSC's petascale *Comet* supercomputer funded by the National Science Foundation (NSF). In the last year, SDSC added 17 new projects in the areas of Pharmacology, Neurological Surgery, Physics, Nanoengineering, Astronomy, EECS, Energy Research, Climate Science, Biology etc. HPC@UC is offered in partnership with the UC Vice Chancellors of Research and campus CIOs. The program has helped UC researchers accelerate their time-to-discovery across a wide range of disciplines, from astrophysics and bioengineering to earth sciences and machine learning. The program is specifically intended to:

- Broaden the base of UC researchers who require advanced and versatile computing;
- Seed promising computational research;
- Facilitate collaborations between SDSC and UC researchers;
- Give UC researchers access to cyberinfrastructure that complements what is available at their campus; and
- Help UC researchers pursue larger allocation requests through the NSF's eXtreme Science and Engineering Discovery Environment (XSEDE) program and other national computing initiatives.

SDSC.EDU/COLLABORATE/HPC_AT_UC.HTML

HIGH-PERFORMANCE COMPUTING AND DATA SCIENCE WORKSHOPS ACROSS THE UC SYSTEM

SDSC staff conducts numerous workshops during the fiscal year to raise awareness among UC researchers about the advantages of using high-performance computing resources such as *Comet* as well as understanding of data science. These sessions, usually just one day, are led by research staff from SDSC's Data-Enabled Scientific Computing (DESC) division and Cyberinfrastructure Research, Education and Development (CI-RED) division. SDSC speakers who teach at these workshops have doctorate degrees in physics, astrophysics, aerospace engineering, computer science, cognitive science, and more. They have attracted several hundred attendees including graduate students, faculty, and postdoctoral researchers at the UCs. Topics covered include machine learning using *Comet*, scaling and optimization of scientific applications in HPC, software tools for life sciences applications, and working with Python and Jupyter notebooks. The workshops, which promote interaction among UC researchers and serve as an easy 'on-ramp' for allocations on *Comet*, have been conducted since the start of the UC@SDSC program in 2014.

SDSC-LED WIFIRE LAB WORKING WITH CALIFORNIA TO COMBAT WILDFIRES

SDSC's WIFIRE Lab is has been working with state local community officials and San Diego Gas & Electric (SDG&E) to share weather and fuel data with the fire science and first response community in a FAIR (Findable, Accessible, Interoperable, and Reusable) and responsible way. In the 2020 fire season, the WIFIRE Lab participated in a statewide public/private project called FIRIS (Fire Integrated Real-Time Intelligence System) led by the Orange County Fire Authority based on the successes of the FIRIS pilot program for Southern California in 2019. Participants in the project include researchers from SDSC and other UC San Diego areas including the California Institute for Telecommunications and Information Technology's (Calit2) Qualcomm Institute, the Mechanical and Aerospace Engineering department at the university's Jacobs School of Engineering, and the Scripps Institution of Oceanography. *(Read more about WIFIRE on page 49)*

UC Researchers Study How to Reset the Biological Clock with a Flip of the Molecular Switch

Scientists from UC San Diego, UC Santa Cruz, and Duke University synchronized their research watches to study what makes our biological clocks tick. They set out to understand why some people are what they call extreme "morning larks" who operate on a shorter 20-hour cycle compared to a regular 24-hour pattern. This difference can result in outcomes such as social jet lag – an inability to function in sync with the rest of society – and sleep disorders. Researchers used the *Triton Shared Computing Cluster (TSCC)* at SDSC to create molecular dynamics (MD) simulations showing how an enzyme called casein kinase 1 (CK1) switches between two conformations, and how mutations cause it to favor one conformation over another. Clock-altering mutations in CK1 had been known for years, but it was unclear how they changed the timing of the clock. "Since circadian rhythms are important for almost every aspect of our physiology, when the biological clock is not properly aligned with Earth's 24-hour day, it can cause a variety of problems ranging from social jet lag to metabolic and degenerative diseases," said UC San Diego's Clarisse Ricci, who co-authored the study, published in the *eLife* science journal.

QRGO.PAGE.LINK/2FFXA

IMPACT & INFLUENCE

CAMPUS AND EDUCATION

Strengthening Ties across Campus and Our Local Communities

SDSC has been expanding its education, outreach, and training (EOT) initiatives at the undergraduate/graduate/postdoctoral levels in research-based computing and data science, specifically artificial intelligence, machine learning, and high-performance computing (HPC) in the cloud. "SDSC's development and operation of HPC resources at the national level provides substantial and tangible benefits to UC San Diego researchers, as well as San Diego's burgeoning research infrastructure," said SDSC Director Michael Norman.

With regard to its national cyberinfrastucture (CI) mission, SDSC had 90 active awards totaling almost $150 million at the end of its 2019/20 fiscal year. Within its campus CI mission, SDSC had about $3.9 million in service agreements, including industry collaborations. On the data science front, SDSC derived about $1.6 million in revenue from its health cyberinfrastructure awards.

EMPOWERING THE NEXT GENERATION OF RESEARCHERS

SDSC's EOT programs start at the high school level to make students aware of opportunities within computational science and engineering at an early age, and then at the university level with numerous data-centric courses, including those in collaboration with the Halicioğlu Data Science Institute (HDSI), based in SDSC's East Building. SDSC also provides range of online courses that attract participants from all over the world. This initiative extends into serving the growing computational science workforce with workshops such as SDSC's Summer Institute, the International Conference on Computational Science, IEEE Women in Data Science Workshop, and more.

SDSC Pilots First Remote GPU Hackathon

The first virtual GPU Hackathon with SDSC successfully concluded in May 2020, marking a new chapter in the evolution of NVIDIA's hackathon program as some 60 participants comprising seven hackathon teams, mentors, and moderators attended. By early March the planning group, including staff from National Energy Research Scientific Computing (NERSC), Oak Ridge National Laboratory (ORNL), SDSC, and other institutes, discussed the possibility of doing the event remotely instead of onsite at SDSC despite myriad potential challenges. "Things were changing daily due to COVID-19," said Susan Rathbun, SDSC's program and events manager. "The planning team introduced a pre-event training session two weeks before the event, which made a big difference in terms of getting staff and attendees up to speed."

ORGO.PAGE.LINK/Zb461

Complementary knowledge graph themes to be integrated into a comprehensive Open Knowledge Network (OKN) by project partners: UC San Diego (blue), UC Santa Barbara (green), and UC San Francisco (orange). Gradients represent areas of overlapping expertise and integration.
Credit: Krzysztof Janowicz, UC Santa Barbara); Peter Rose and Ilya Zaslavsky, SDSC/UC San Diego.

On the Local Front...

SDSC RESEARCHERS CREATE COVID-19 KNOWLEDGE GRAPHS FOR SAN DIEGO REGION

In May 2020 the National Science Foundation (NSF) awarded two SDSC researchers funding to organize COVID-19 information into a transdisciplinary knowledge network that integrates health, pathogen, and environmental data to better track cases to improve analysis and forecasting across the greater San Diego region.

The six-month award allowed the researchers to quickly launch a comprehensive semantic integration platform for data-driven analysis and development of policy interventions, and took into account social-economic, and demographic characteristics of populations in different areas as well as biomedical data such as virus strains and genetic profiles, said Peter Rose, director of SDSC's Structural Bioinformatics Laboratory and principal investigator for the project. *(Read more about this UC-wide collaboration on page 5 and SDSC's role in life science on page 40)*

ORGO.PAGE.LINK/YU9SM

Background credit: Erik Jepsen, UC San Diego Publications.

Credit: UC San Diego CREATE

Robert Sinkovits is director of SDSC's Education and Training program.

UC SAN DIEGO'S CREATE AND SDSC AWARDED NATIONAL K-12 STEM GRANT

In mid-2020 the U.S. Department of Defense's (DoD) Defense STEM Education Consortium (DSEC) awarded a one-year grant to SDSC and the UC San Diego Mathematics Project housed at UC San Diego's Center for Research on Educational Equity, Assessment, and Teaching Excellence (CREATE), to introduce computing into high school math classrooms.

The DSEC Innovation Bloc grant, called ICAT through DM - short for Introducing Computing and Technology through Discrete Math problem-solving – funded a Summer Institute for regional high school mathematics teachers and a Summer Academy for military-connected, underserved, and underrepresented rising seniors.

Robert Sinkovits, director of SDSC's scientific computing applications and lead for SDSC's Education and Training program; and Osvaldo Soto, director of the UC San Diego Mathematics Project at CREATE, are the principal investigators for the grant.

The DSEC Innovation Bloc grant was awarded as part of a five-year, $75 million grant from the DoD under which DSEC focuses on K-16 (kindergarten through college) STEM enrichment programs for military-connected and/or low-income students and educators, as well as workforce engagement, program evaluation, and public outreach efforts across the nation. The consortium is comprised of 18 organizations, and includes UC San Diego CREATE, which serves as a program hub lead on the DSEC grant.

QRGO.PAGE.LINK/NNP6N

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 (SC) conference. 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 PIs to create new CCRs. In late 2020 SDSC successfully fielded its first team for the annual Student Cluster Competition (SCC) at the annual Supercomputing Conference. SCC was started in 2007 to provide an immersive HPC experience for undergraduate and high school students.

BIT.LY/SDSC_HPC_STUDENTS

EDUCATIONAL IMPACT&INFLUENCE
ACROSS CAMPUS AND OUR LOCAL COMMUNITIES

High school students participated in the REHS Project Showcase where they shared research projects with peers, mentors, family, and friends.

Ange Mason is SDSC's education manager and the founder of REHS.

FIRST VIRTUAL RESEARCH EXPERIENCE FOR HIGH SCHOOL STUDENTS

SDSC's Research Experience for High School Students (REHS) program, which celebrated its 11th year in 2020, was developed to help increase awareness of computational science and related fields of research among students in the greater San Diego area. The eight-week program – conducted remotely for the first time due to the COVID-19 pandemic – paired 54 students with 12 SDSC mentors to help them gain experience in an array of computational research areas, while gaining exposure to career options and work readiness skills. Capping off 2020'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 460 students have participated in SDSC's REHS program.

"Despite some apprehension among students about the program being conducted remotely because of the pandemic, we found that they quickly adapted to the virtual format and enjoyed what was an educationally enriching experience," said SDSC Education Manager and REHS Founder Ange Mason. "The majority of students felt they were not significantly impacted by the online environment, but many still let us know that they missed the in-person interaction with both their mentors and other interns."

QRGO.PAGE.LINK/GLVVX

MENTOR ASSISTANCE PROGRAM

San Diego-area high school students interested in pursuing a career in scientifically-based 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 UC San Diego School of Medicine, MAP's mission is to provide a pathway for student researchers to gain access to UC San Diego faculty, postdoctoral fellows, Ph.D. candidates, and staff to mentor them in their own field of interest. Mentors are recruited from across campus from fields that include athletics, biology, chemistry, aerospace engineering, network architectures, pharmaceutical sciences, physics, social studies, and more.

QRGO.PAGE.LINK/UVFCV

Background credit: Erik Jepsen, UC San Diego Publications.

EDUCATIONAL IMPACT&INFLUENCE
On the National Front

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 campus' 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 in every populated continent – from Uruguay to the Ivory Coast to Bangladesh. A subset of students pays for a certificate of completion.

ORGO.PAGELINK/ENYMG

HPC TRAINING WEBINARS & WORKSHOPS

SDSC's High-Performance Computing (HPC) training webinar and workshop series focus on familiarizing researchers who would like to learn more about using HPC systems (*Comet, Expanse, Triton Shared Computing Cluster*) resources. Through mid-2020 there have been 36 training events with over 3,700 participants from both UC San Diego as well as nationally via the XSEDE program. Topics include 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. Members of the HPC Training team began working with Advanced Micro Devices (AMD) to form the AMD HPC User Forum to start in late 2020.

BIT.LY/SDSC_HPC_TRAINING

HIGH-PERFORMANCE COMPUTING SUMMER INSTITUTE

SDSC's annual week-long training program offering introductory to intermediate topics on high-performance computing (HPC) and data science with hands-on tutorials using SDSC's *Comet* supercomputer was held virtually due to the COVID-19 pandemic. The program covers machine learning at scale, distributed programming in Python and data management, and more traditional HPC topics such as performance tuning, CUDA programming, and visualization and parallel programming with MPI and OpenMP. The 2020 institute had 55 participants from 37 institutions and companies, spanning four time zones. To avoid "Zoom fatigue", the length of the days was reduced and participants were required to attend one of two tech checks so that the event could remain fully dedicated to instruction. Graduate students made up about half of the attendees, while about one-quarter were staff and the remainder postdoctoral researchers, faculty, librarians, and industrial users. SDSC has conducted the Summer Institute since the mid-1990s.

SI20.SDSC.EDU

DATA EAST CONFERENCE

The inaugural Data East technology conference, held in August 2020 in a virtual format, was attended by some 100 invitation-only participants. Its theme was 'Innovating through Crisis: Defining Technology-Enabled Ecosystems of the Future.' Keynote speakers included Jamie Holcombe, chief information officer for the U.S. Patent and Trademark Office; Professor Lynda Applegate of Harvard Business School; Steve Orrin, chief technology officer at Intel Federal; Dean of Engineering and Professor Maj Mirmirani of Embry-Riddle Aeronautical University; and IBM Vice President Naguib Attia. The event, chaired by Applegate and SDSC Lead Scientist James Short, featured two breakout session tracks. Speakers in the ecosystem policy/governance track discussed two unique public-private partnerships developing in Chesapeake Bay and the city of Detroit. Speakers in the new technologies track covered current projects in SDSC's BlockLAB, including a presentation by Scott Kahn of LunaPBC on healthcare data privacy.

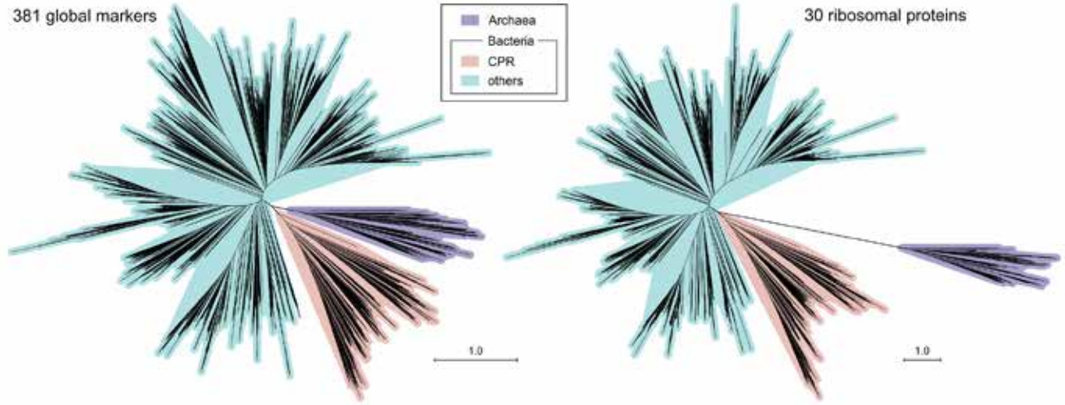
SCIENCE HIGHLIGHTS

HUMAN HEALTH | LIFE SCIENCES

From Combating Viruses to Better Understanding the Tree of Life

SDSC Supercomputers Advance Discovery

Please read about SDSC's resource support and research initiatives related to COVID-19 on page 4



A multinational study led by UC San Diego researchers found that the evolutionary distance between Archaea and Bacteria domains is much less in the left tree obtained from a global set of 381 global marker genes than in the right tree obtained from only 30 genes for ribosomal proteins, similar to those used in other studies. This study provides strong evidence that trees based on more broadly selected genes better reflect genome-level evolution and a more accurate view of the tree of life. *Credit: Qiyun Zhu, et al.*

Study Finds Close Evolutionary Proximity Between Microbial Domains in the 'Tree of Life'

A comprehensive analysis of 10,575 genomes as part of a multinational study led by UC San Diego researchers reveals close evolutionary proximity between the microbial domains at the base of the tree of life, the branching pattern of evolution described by Charles Darwin more than 160 years ago. The study, published in *Nature Communications* in December 2019, found much closer evolutionary proximity between the Archaea and Bacteria microbial domains than have most previous studies. This new result arises from the use of a comprehensive set of 381 marker genes versus a couple of dozen core genes such as ribosomal proteins typically used in previous studies, according to Qiyun Zhu, a postdoctoral scholar in the UC San Diego School of Medicine's Department of Pediatrics and lead author of the paper. SDSC Distinguished Scientist Wayne Pfeiffer made more than 2,000 runs on the standard compute nodes of *Comet* to generate the gene trees, while Uyen Mai, a Ph.D. student in the Mirarab Lab at UC San Diego and co-first author of the paper, combined these trees using ASTRAL on *Comet's* GPU nodes.

QRGO.PAGE.LINK/WS9FC

Unlocking Reproductive Mysteries of Viruses and Life

One such study used supercomputer simulations conducted on *Comet* to determine a chemical mechanism for the reaction of nucleotide addition, used in the cell to add nucleotide bases to a growing strand of DNA. "Researchers were also able to determine the role of a catalytic metal ion of magnesium that's in the active site of the enzyme DNA polymerase," said study co-author Daniel Roston, an assistant project scientist in the Department of Chemistry and Biology at UC San Diego. "This metal has been a bit controversial in the literature. Nobody was really sure exactly what it was doing there. We think it's playing an important catalytic role."

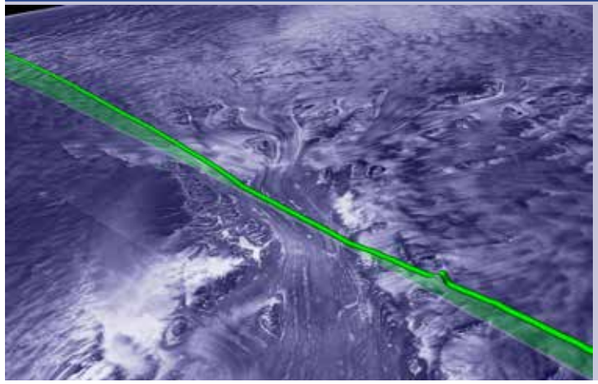
The chemistry needs multiple proton transfers in a complex active site. Experimental probes using X-ray crystallography have been unable to distinguish among the many possible reaction pathways. "Such simulations offer a compliment to crystallography because one can model in all the hydrogens and run molecular dynamics simulations, and allow the atoms to move around in the simulation and see where they want to go, and what interactions are helping them," said Roston. He and his colleagues used some 500,000 CPU hours on *Comet*, which enabled them to simultaneously run many different simulations that fed off one another. The study was published in the December 2019 *Proceedings of the National Academy of Sciences*.

QRGO.PAGE.LINK/G9QHK

EARTH SCIENCES

Understanding Fire & Ice through the Power of Data & Computation

SCIENCE HIGHLIGHTS



ICESat's topographic profiles across the continent reveal the textured surfaces of Antarctic ice sheets in unprecedented detail. At left, a slice of (green) elevation data passes over Lambert Glacier.

Credit: NASA/Goddard Space Flight Center Scientific Visualization Studio, Canadian Space Agency, RADARSAT International Inc.

NASA Extends Support for OpenAltimetry Data Platform

In mid-2020 NASA extended support for OpenAltimetry, a web-based cyberinfrastructure platform that enables discovery, access, and visualization of altimetry data from NASA's ICESat (Ice, Cloud, and land Elevation Satellite) and ICESat-2 (launched in November 2018) laser altimeter missions. These laser profiling altimeters are being used to measure changes in the topography of Earth's ice sheets, vegetation canopy structure, and clouds and aerosols.

Initially funded by a NASA ACCESS (Advancing Collaborative Connections for Earth System Science) grant in 2016, OpenAltimetry is a collaborative project between SDSC, Scripps Institution of Oceanography, National Snow and Ice Data Center, and UNAVCO.

OpenAltimetry facilitates a new paradigm for access to these NASA mission datasets to serve the needs of a diverse scientific community as well as increase the accessibility and utility of data for new users. NASA's support enables OpenAltimetry to continue providing access to ICESat-2 data products. These datasets will continue to expand as ICESat-2 collects new data in the coming years.

OPENALTIMETRY.ORG

Since early 2019, OpenAltimetry has enabled over 15,250 ICESat-2 data downloads via the portal and some 540,000 downloads via its API, as well as visualization of almost 137,000 elevation and 93,000 photon plots.

Creating Simulations for Tsunami Case Study

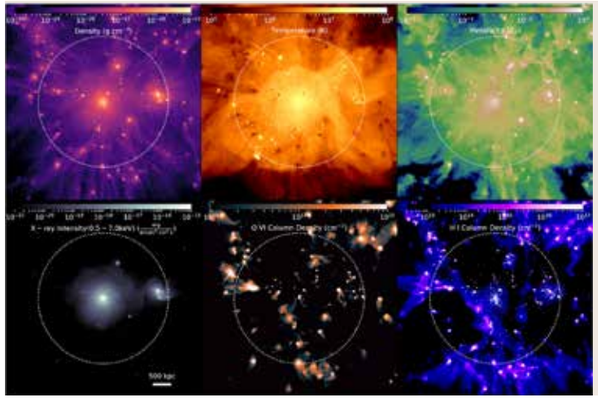
Researchers at the University of Rhode Island (URI) used SDSC's *Comet* supercomputer to show that high-performance computer modeling can accurately simulate tsunamis from volcanic events. Such models could lead to early-warning systems that could save lives and help minimize catastrophic property damage.

URI Distinguished Professor and Chair of the Department of Ocean Engineering Stephan Grilli and his team published their findings in the *Nature Scientific Reports*. Their paper focused on the December 22, 2018 collapse of the Anak Krakatau volcano and subsequent tsunami, which was the first time in recent history an event such as this happened. The event allowed researchers an opportunity to test their models and modeling methodologies for accuracy against the recorded observations.

The URI scientists created their simulations while working with British Geological Survey collaborators, who surveyed the Indonesian area several times in 2019, where more than 400 people were killed during the late 2018 event. Their simulations successfully reproduced post-tsunami characteristics, tide gauge records, and eyewitness reports – suggesting that their landslide volume range and assumed collapse scenarios were accurate.

QRGO.PAGE.LINK/2BWJY

Ultimately, their supercomputer simulations demonstrated that, in cases such as Anak Krakatau, the absence of precursory warning signals, together with the short travel time following tsunami initiation, presents a major challenge for mitigating tsunami coastal impact. "We're hopeful that our continued research reduces warning systems from several hours to approximately 10 minutes so that more people can reach safety prior to a tsunami," said Grilli.



A 5x5 megaparsec (~18.15 light years) snapshot of the RomulusC simulation at redshift z = 0.31. The top row shows density-weighted projections of gas density, temperature, and metallicity. The bottom row shows the integrated X-ray intensity, O VI column density, and H I column density.

Credit: Iryna Butsky et al.

Supercomputer Simulations Reveal Secrets of the Universe

Black Holes and Galaxy Clusters

Inspired by the Romulans, a fictional extraterrestrial race in the 'Star Trek' series, astrophysicists have developed cosmological computer simulations called RomulusC, where the 'C' stands for galaxy cluster. With a focus on black hole physics, RomulusC has produced some of the highest resolution simulations ever of galaxy clusters, which can contain hundreds or even thousands of galaxies.

On Star Trek, the Romulans powered their spaceships with an artificial black hole. In reality, it turns out that black holes can drive the formation of stars and the evolution of whole galaxies. In short, scientists were able to probe the intracluster medium which fills the space between galaxies in a galaxy cluster but is also invisible to optical telescopes.

QRGO.PAGE.LINK/YXUXQ

For their study, published in the *Monthly Notices of the Royal Astronomical Society*, the researchers used several supercomputers, including SDSC's *Comet* system, which fills a particular niche, according to study co-author Tom Quinn, a professor of astronomy at the University of Washington. "Having a large shared memory machine was very beneficial for particular aspects of the analysis, for example identifying the galaxies, which is not easily done on a distributed-memory machine."

Novel Planet Formation Models

Most of us are taught in grade school how planets formed: dust particles clump together and over millions of years continue to collide until one is formed. This lengthy and complicated process was recently modeled using a novel approach with the help of SDSC's *Comet* supercomputer by scientists at the Southwest Research Institute (SwRI), who created a simulation of planet formation that provides a new baseline for future studies of this mysterious field.

"Specifically, we modeled the formation of terrestrial planets such as Mercury, Venus, Earth, and Mars," said Kevin Walsh, SwRI researcher and lead author of a study of published in the *Icarus Journal*. "The problem of planet formation is to start with a huge amount of very small dust that interacts on super-short timescales (seconds or less), and the *Comet*-enabled simulations finish with the final big collisions between planets that continue for 100 million years or more."

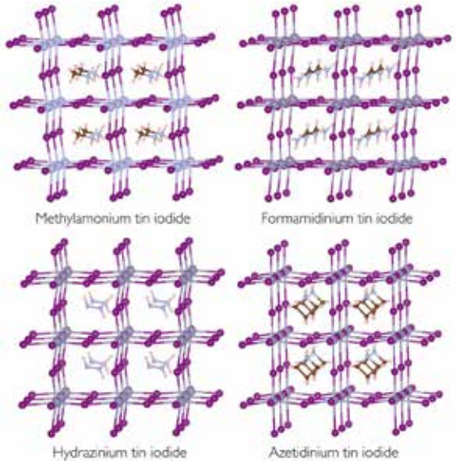
These models give us insight into the key physics and timescales involved in our own solar system, according to the researchers. They also allow us to better understand how common planets such as ours could be in other solar systems – meaning that environments similar to Earth may exist. "Part of this puzzle is to understand how the ingredients of life, such as water, made their way to Earth," said Walsh. "One big consideration is these models traced the material in the solar system that we know is rich with water, and seeing what important mechanisms can bring those to Earth and where they would have done so."

QRGO.PAGE.LINK/JUQ95

MATERIALS ENGINEERING

From Energy Savings to Improving Our Environment

Supercomputer Simulations Reveal New Possibilities



Four lead-free perovskites were simulated using SDSC's *Comet* supercomputer and *Stampede2* at the Texas Advanced Computing Center. These simulations show that these materials exhibit promising features for solar energy options. They are now being synthesized for further investigation.

Credit: H. Tran et al (Georgia Institute of Technology), V. Ngoc Tuoc (Hanoi University of Science and Technology)

Shedding Light on Inexpensive, Efficient Solar Energy

Solar energy has become a popular renewable source of electricity around the world with silicon serving as the primary source due to its efficiency and stability. Because of silicon's relatively high cost, hybrid organic-inorganic perovskites (HOIPs) have emerged as a lower-cost – and highly efficient – option for solar power, according to a recent study by Georgia Institute of Technology researchers who used SDSC's *Comet* supercomputer and *Stampede2* at the Texas Advanced Computing Center.

The name perovskite refers not only to a specific mineral ($CaTiO_3$) found in Russia's Ural Mountains, but also to any compound that shares its structure. A search for stable, efficient, and environmentally safe perovskites has created an active avenue in current materials research. However, the presence of lead in the most promising perovskite candidates, methylammonium and formamidinium lead halides, has raised concerns. Moreover, these materials have shown to be unstable under certain environmental conditions.

The Georgia Tech researchers worked with colleagues at the Hanoi University of Science and Technology in Vietnam to create simulations that identified four lead-free perovskites as promising candidates for solar cell materials. Two of them have already been synthesized and the other two are recommended for further investigations.

QRGO.PAGE.LINK/UMGX8

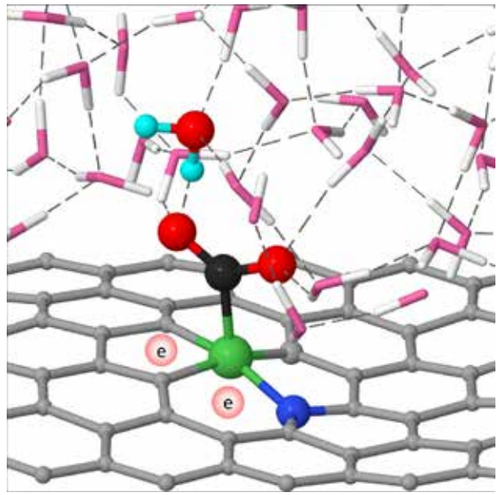
The research, which won a 2020 HPCWire award for Top Energy-Efficient HPC Achievement, was published in *The Journal of Chemical Physics*.

Revealing the True Strengths of Zirconia Ceramics

For thousands of years, humans have produced ceramics by simply combining specific minerals with water or other solvents to create ceramic slurries that cure at room temperature and become some of the hardest known materials. In more recent times, zirconia-based ceramics have been useful for an array of applications ranging from dental implants and artificial joints to jet engine parts.

Researchers from the Colorado School of Mines have been using multiple supercomputers, including SDSC's petascale *Comet* system, to study certain characteristics of zirconia. Their findings in the *Journal of the European Ceramic Society* focused on assessing the ability of zirconia-based ceramics to withstand harsh conditions as well as the extreme limits of fracture and fatigue. "By using large-scale atomistic simulations, these simulations revealed how specific nanoscale structures, twin boundaries, and pre-existing defects control the mechanical behavior and the corresponding plastic deformation of an advanced shape memory ceramic called yttria-stabilized tetragonal zirconia or YSTZ," said corresponding author Mohsen Asle Zaeem, a mechanical engineering professor at the Colorado School of Mines.

QRGO.PAGE.LINK/DZSMX



Electrochemical reduction of carbon dioxide catalyzed by single nickel atom embedded in graphene with nitrogen dopant;"e" represents the electrons.

Credit: Xunhua Zhao and Yuanyue Liu, UT-Austin

Supercomputer Simulations Help Advance Electrochemical Reaction Research

Single-atom catalysts have emerged as promising innovations for solving environmental and energy issues. One such example, nickel embedded in graphene (a thin layer of graphite), has been shown to convert carbon dioxide, a molecule that causes the greenhouse effect, into carbon monoxide, an important feedstock for chemical engineering.

As a poisonous gas, carbon monoxide is often converted into carbon dioxide, such as in cars and trucks equipped with catalytic converters. This process is the reverse, which at first may sound a bit odd, but provides an important role in synthesizing valuable chemicals to use as electricity in lieu of the Earth's quickly depleting fossil fuels. However, a better understanding of the atomic structure of this concoction is needed before nickel-embedded graphene can be used on a regular basis.

To help with this challenge, researchers from the University of Texas at Austin (UT-Austin) simulated the catalytic mechanism and atomic structure of nickel-doped graphene using SDSC's *Comet* and *Stampede2* at the Texas Advanced Computing Center (TACC). The simulations, published in the *Journal of The American Chemical Society*, showed a clear picture of the catalyst's atomic structure so that researchers were able to better understand critical effects of surface change and hydrogen bonding, which were overlooked in previous models.

QRGO.PAGE.LINK/EZQJC

SDSC Awarded HPCwire Editors' Choice for Top Energy-Efficient HPC Achievement

In late 2019, SDSC received three top HPCwire awards, including Editors' Choice for the Top Energy-Efficient HPC Achievement. The award recognized UC San Diego researchers for using *Comet* to design new materials for solar cells and LED's, anticipating that these materials will provide excellent properties for this application. The awards were presented at the the International Conference for High-Performance Computing, Networking, Storage, and Analysis (SC19) in Denver, CO.

QRGO.PAGE.LINK/ZOWVB

FOCUSED SOLUTIONS & APPLICATIONS

FOR **ADVANCED COMPUTATION** AND **RESEARCH DATA SERVICES**

SDSC's expertise in high-performance computers, data storage, and networking has helped create an advanced cyberinfrastructure that supports and accelerates scientific discovery across academia, industry, and government.

Advanced, feature-rich resources such as SDSC's new *Expanse* supercomputer help meet the need for systems that can serve a broad range of science domains while reducing the barriers to understanding and using such systems. As one of the nation's top academic supercomputing centers, SDSC has focused on researchers who have modest to medium-scale computational needs – which is where the bulk of computational science needs exist.

*The **"long tail" of science** is the idea that the large number of modest-sized computationally based research projects represent, in aggregate, a tremendous amount of research that can yield scientific advances and discovery.*

Expanse

Computing Without Boundaries

In late 2020 SDSC launched its newest National Science Foundation (NSF)-funded supercomputer, *Expanse*. At over twice the performance of *Comet*, *Expanse* supports SDSC's theme of 'Computing without Boundaries' with a data-centric architecture and state-of-the art GPUs for incorporating experimental facilities and edge computing.

Like *Comet*, *Expanse* is suited for modest-scale jobs as few as tens of cores to several hundred cores, and can also handle high-throughput computing jobs via integration with the Open Science Grid, which can have tens of thousands of single-core jobs. *Expanse* also provides connectivity to commercial clouds via the job queuing system. A low-latency interconnect based on Mellanox High Data Rate (HDR) InfiniBand supports a fabric topology optimized for jobs of one to a few thousand cores that require medium-scale parallelism.

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 entire system, integrated by Dell, is organized into 13 SDSC Scalable Compute Units (SSCUs), comprising 56 standard nodes and four GPU nodes, and connected with 100 GB/s HDR InfiniBand. Direct liquid cooling to the compute nodes provides high core count processors with a cooling solution that improves system reliability and contributes to SDSC's energy-efficient data center.

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. The *Expanse* cluster is managed using the Bright Computing HPC Cluster management system, and the SLURM workload manager for job scheduling. Local NVMe on each node gives users a fast scratch file system that dramatically improves I/O performance of many applications. In 2021, a Ceph-based file system will be added to *Expanse* to support complex workflows, data sharing, and staging to/from external sources.

A key innovation of *Expanse* is its ability to support composable systems, which can be described as the integration of computing elements (i.e., some number of compute elements, GPU, large memory nodes) into scientific workflows that may include data acquisition and processing, machine learning, and traditional simulation. *Expanse* also has direct scheduler-integration with the major cloud providers, leveraging high-speed networks to ease data movement to/from the cloud.

Like *Comet*, *Expanse* is a key resource within the NSF's Extreme Science and Engineering Discovery Environment (XSEDE), which comprises the most advanced collection of integrated digital resources and services in the world. The NSF award for *Expanse* runs from October 1, 2020 to September 30, 2025, and is valued at $10 million for acquisition and deployment of *Expanse*. An additional award will support *Expanse* operations and user support.

EXPANSE.SDSC.EDU

COMET

A COMPUTING CYBERINFRASTRUCTURE FOR THE 'LONG TAIL' OF SCIENCE

SDSC's computational, storage, and networking resources – plus a high level of combined expertise required to configure, operate, and support them – create an advanced cyberinfrastructure that supports scientific discovery among numerous disciplines across academia, industry, and government.

Advanced but user-friendly resources such as SDSC's petascale-level *Comet* supercomputer underscore a vital need for systems that serve a broad range of research, with a focus on researchers who have modest to medium-scale computational needs, which is where the bulk of computational science needs exist.

While *Comet* is capable of an overall peak performance of 2.76 petaflops – or 2.76 quadrillion calculations per second – its allocation and operational policies are geared toward rapid access, quick turnaround, and an overall focus on scientific productivity.

"SDSC's national mission to help pioneer an advanced research cyberinfrastructure has always been our core, and that has enabled us to support collaborations at the local and state levels," said SDSC Director Michael Norman, principal investigator (PI) for the *Comet* program, the result of NSF grants now totaling more than $27 million.

100,000*
Individual users have run jobs on *Comet* through Science Gateways**

1,600,000,000+*
Core hours of computing provided by *Comet* since it began operations

8,500*
Individual users accessed *Comet* via traditional login from over 400 unique institutions

10,000,000+*
GPU hours of computing provided by *Comet* since it began operations

In mid-2018 the NSF extended *Comet's* service into a sixth year of operation, with the system now slated to run through July 31, 2021.

*As of mid-2020. ** A science gateway is a community-developed set of tools, applications, and data services and collections that are integrated through a web-based portal or suite of applications.

VOYAGER

ADVANCING ARTIFICIAL INTELLIGENCE

In mid-2019 the NSF awarded SDSC a $5 million grant to develop a high-performance resource for conducting artificial intelligence (AI) research across a wide swath of science and engineering domains. The experimental system, to be called *Voyager*, will be the first-of-its-kind available in the NSF resource portfolio. The NSF award is structured as a three-year 'test bed' phase to start in mid-2021, followed by a two-year phase beginning in mid-2024; in which allocations will be made using an NSF-approved process.

Using AI processors optimized for deep learning operations, *Voyager* will provide an opportunity for researchers to explore and evaluate the system's unique architecture using well-established deep learning frameworks to implement deep learning techniques. Researchers will also be able to develop their own AI techniques using software tools and libraries built specifically for *Voyager*.

In advance of the *Voyager* award, SDSC established the AI Technology Lab (AITL) in October 2019, which provides a framework for leveraging *Voyager* to foster new industry collaborations aimed at exploring emerging AI and machine learning technology for scientific and industrial uses, while helping to prepare the next-generation workforce.

"A rigorous evaluation of hardware has been optimized for AI algorithms of keen interest across the entire AI research community," said Amitava Majumdar, head of SDSC's Data Enabled Scientific Computing division and PI for the *Voyager* project. Co-PIs include Rommie Amaro, a professor of chemistry and biochemistry and director of the National Biomedical Computation Resource at UC San Diego; and UC San Diego Physics Professor Javier Duarte. SDSC co-PIs are Robert Sinkovits, lead for scientific applications and Mai Nguyen, lead for data analytics.

COMET
QRGO.PAGE.LINK/2NBW2

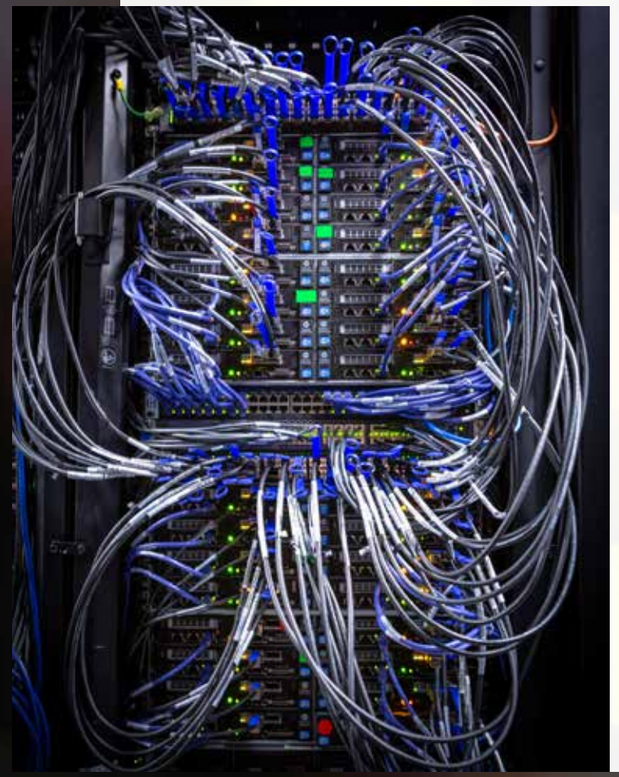
VOYAGER PRESS RELEASE
QRGO.PAGE.LINK/9ZZCW

TRITON SHARED COMPUTING CLUSTER

The *Triton Shared Computing Cluster (TSCC)* is a medium-scale high-performance computing cluster operated on behalf of UC San Diego by SDSC. Focusing primarily on campus investigators, *TSCC* supports research across a wide variety of domains, including biology and life sciences, chemistry, climate science, engineering, political and social sciences, and more. *TSCC* also serves several other UC campuses and private sector researchers in areas such as genomics and materials science. Launched in 2013, *TSCC* arose out of a need for campus researchers to have ready access to a local high-performance computing resource as modeling, simulations, and data analytics experienced growing use in virtually every avenue of scientific investigation. *TSCC* comprises almost 300 "condo" (researcher-purchased) nodes (49 condo groups) and 37 "hotel" (SDSC-purchased) nodes, the latter of which are available to users through a pay-as-you-go recharge model.

During mid-2019 to mid-2020, *TSCC* underwent major modernizations to address the evolving needs of scientific computing. The *TSCC* hotel partition was refreshed with updated compute nodes, providing more computational cores with higher performance in a reduced physical node count for better efficiency. The *TSCC* condo group welcomed the participation of the Scripps Institute of Oceanography (SIO), with a group of researchers migrating from an aging departmental cluster. The SIO group purchased Dell high-density compute nodes with the latest Intel "Cascade Lake" processors to meet their diverse computing needs. Work proceeded in 2019-2020 on a National Science Foundation (NSF) Campus Cyberinfrastructure (CC*) grant called Triton Stratus. The grant runs until mid-2021 and includes funding for new on-campus capabilities as well as credits for commercial cloud computing services.

GOO.GL/WWAQL5



Triton Stratus provides researchers with improved facilities for utilizing emerging computing paradigms and tools, namely interactive and portal-based computing, and scaling them to commercial cloud computing resources. This capability accommodates the growing need of researchers, especially data scientists, that use tools such as Jupyter notebooks and RStudio® to implement computational and data analysis functions and workflows. The project is also investigating and deploying techniques for packaging tested notebooks and sending or "bursting" them to commercial cloud services for greater scale or throughput.

SDSC Director of Industry Relations Ron Hawkins is the principal investigator for Triton Stratus, joined by SDSC's Robert Sinkovits, Subhashini Sivagnanam, and Mary Thomas as co-PIs. The *TSCC* group also began laying the foundation for an improved cluster management framework, including Bright Cluster Manager and the Slurm cluster scheduling system, which will provide a more stable and maintainable software environment. In total, these modernization efforts are positioning *TSCC* to continue to serve as a valuable resource to meet the growing needs of campus researchers for scientific computing support.

SDSC COLOCATION FACILITY

SDSC offers colocation (colo) services to UC San Diego, the UC system, and the local research community. SDSC's 19,000 square-foot, climate-controlled, secure data center is designed for maximum power efficiency and power density, multi-40, and 100 gigabit network connectivity, and a 24/7 operations staff. This capability has been essential during the COVID-19 pandemic as a way for most staff and research teams to work from home. SDSC Operations staff provide remote hands service to all Colo customers. Within that center is a "colocation" facility that is free to UC San Diego researchers via a program aimed at saving campus funds by housing equipment in an energy-efficient, shared facility.

The SDSC Colo has a special compliance zone for customers storing HIPAA, FISMA, and other sensitive data. The facility houses computing and networking equipment for dozens of campus departments, every division and school, as well as local partners that include Rady Children's Hospital, the J. Craig Venter Institute, Simons Foundation, The Scripps Research Institute, and the Sanford-Burnham Medical Research Institute. SDSC's Colo facility has resulted in more than $2.1 million in annual energy savings, while streamlining and improving the management of hundreds of campus systems. The facility is well-suited to installations that need to demonstrate regulatory compliance, as well those that require high-speed networking. SDSC welcomes inquiries from local companies interested in co-locating equipment to facilitate collaborations with UC San Diego and SDSC investigators.

STORAGE, NETWORKING, AND CONNECTIVITY

SDSC's Research Data Services team administers a large-scale storage and compute cloud. UC San Diego campus users, members of the UC community, and UC affiliates are eligible to join the hundreds of users who already benefit from the multi-petabyte, OpenStack Swift object store. SDSC Cloud has a simplified recharge plan that eliminates fees such as bandwidth and egress charges. SDSC Cloud also includes an elastic compute facility, based on OpenStack Nova, using Ceph for storage. This comprehensive cloud environment provides researchers with a testbed and development environment for developing cloud-based services, and for many data science workflows. It is especially attuned to data sharing, data management, and data analytics services. The UCSD campus encourages research faculty to make use of shared systems through an indirect cost (IDC) waiver for SDSC Cloud and other shared services. *(See also CloudBank and USS on pages and 9 and 11)*

QRGO.PAGE.LINK/WNWDP

FOCUSED SOLUTIONS & APPLICATIONS

FOR LIFE SCIENCES COMPUTING

Life sciences computing has been a key element of SDSC's strategic plan, with the goal of improving the performance of bioinformatics applications and related analyses using the Center's advanced computing systems. The initial work, co-sponsored and supported by Dell and Intel, involved benchmarking selected genomic and Cryo-electron Microscopy (Cryo-EM) analysis pipelines. SDSC's initiative focuses on developing and applying rigorous approaches to assessing and characterizing computational methods and pipelines.

SDSC Recognized in HPCwire Editors' Choice Award for Best Use of HPC in the Life Sciences

In late 2019 online publication HPCwire awarded SDSC and researchers at the University of Michigan (UM) for using the Center's *Comet* supercomputer to analyze differences between 2D and 3D visualizations to understand how tuberculosis granulomas form and spread. SDSC won this category in 2018 for assisting in research related to autism spectrum disorder (ASD). "For more than 15 years, computer (in silico) modeling has been used to provide insight to the lethal disease," said Simeone Marino, an associate research scientist at UM's Medical School and first author in the recently published study. The computational tool used to analyze the complex dynamics of TB granuloma formation and progression is called agent-based model (ABM) and referred to as GranSim, for granuloma simulation. It works like a virtual reality environment, where cells, bacteria, and molecules are represented by various agents of different pixel sizes in an abstract 3D cube. The cube captures a section of the lung where granulomas typically form.

QRGO.PAGE.LINK/TYIRA

Open EEGLAB Portal Advances UC San Diego's Neuroscience Gateway Project

Even though electroencephalography (EEG) has been used for almost 100 years, this safe and painless test of brain activity remains an efficient method for recording aspects of rapid brain activity patterns supporting our thoughts and actions. EEGLAB is the mostly widely used Matlab based EEG data processing tool and has been developed by Arnaud Delorme, Ramon Martinez, and Scott Makeig of UC San Diego's Swartz Center for Computational Neuroscience. In the last few years EEGLAB has been made available via the Neuroscience Gateway (NSG) on SDSC's *Comet* supercomputer in collaboration with SDSC researchers Amitava Majumdar, Subhashini Sivagnanam, and Kenneth Yoshimoto. In 2020 an EEGLAB plug-in was released and interfaces EEGLAB with NSG directly from within EEGLAB running on MATLAB on any personal lab computer. The plug-in features a flexible MATLAB graphical user interface that allows users to easily submit, interact with, and manage NSG jobs, and to retrieve and examine their results. This will further enable and make it easy for cognitive neuroscientists to process large amounts of EEG data for which there is an increasing need for using high-performance computing (HPC) resources.

QRGO.PAGE.LINK/NVIFK

Reproducible and Scalable Structural Bioinformatics: Application to COVID-19

Scientists face time-consuming barriers when applying structural bioinformatics analysis, including complex software setups, non-interoperable data formats, and ever-larger data sets that need to be analyzed, all which make it difficult to reproduce results and reuse software pipelines. To address these issues, SDSC's Structural Bioinformatics Laboratory, directed by Peter Rose, is developing a suite of reusable, scalable software components called 'mmtf-pyspark', built on the Apache Spark analytics engine for large-scale parallel data processing. Applications developed on top of this platform include 'mmtf-proteomics' to map post-translational modifications, and 'mmtf-genomics' to map mutations to 3D protein structures and evaluate the effect on disrupting protein-protein interactions and drug molecule binding. *(Please see page 4 for more life science research related to SDSC's contribution to COVID-19 research)*

Measuring Mutations in Sperm May Reveal Risk for Autism in Future Children

While the causes of autism spectrum disorder (ASD) are not fully understood, researchers believe both genetics and environment play a role. In some cases, the disorder is linked to *de novo* mutations that appear only in the child and are not inherited from either parent's DNA. In a recent study published in *Nature Medicine*, an international team of scientists led by researchers at UC San Diego's School of Medicine described a method to measure disease-causing mutations found only in the sperm of the father, providing a more accurate assessment of ASD risk in future children. The research team used SDSC's *Comet* to align the whole genome sequences.

"Autism afflicts one in 59 children and we know that a significant portion is caused by these *de novo* DNA mutations, yet we are still blind to when and where these mutations will occur," said co-senior author Jonathan Sebat, professor and chief of the Beyster Center for Molecular Genomics of Neuropsychiatric Diseases at the UC San Diego School of Medicine. "With our new study, we can trace some of these mutations back to the father, and we can directly assess the risk of these same mutations occurring again in future children."

QRGO.PAGE.LINK/NDSW7

FOCUSED SOLUTIONS & APPLICATIONS

FOR **DATA-DRIVEN PLATFORMS** AND **APPLICATIONS**

SDSC's mission has steadily expanded to address more than advanced computation as researchers require innovative applications that address the ever-increasing amount of digitally based scientific data. Within the last 18 months or so the Center has made significant inroads in cloud-based computing, artificial intelligence, and machine learning.

"SDSC's expertise in scalable scientific computing is vital to supporting the increase in data-enabled scientific research," said SDSC Director Michael Norman. "Our newest supercomputer, *Expanse*, which went into service in late 2020, includes cloud integration that gives users access to the latest GPUs and processors as they become available from public cloud providers and computer technology companies."

DATASCIENCE.SDSC.EDU

The Data Science Hub at SDSC

A key goal of the Data Science Hub (DSH) at SDSC is to foster partnerships across academia, industry, and government while serving as a hub of connectivity and collaboration for projects requiring high-level data analytics, 'big data' management, and distributed computing expertise, starting with cross-campus collaborations. "In order to effectively respond to today's data-rich research priorities, there must be a fertile ground of communication and collaboration between like-minded researchers across campus, the UC system, and the active data science startup community in San Diego," said SDSC Chief Data Science Officer Ilkay Altintas.

Another area of focus for the DSH is education and outreach efforts, including the development of professional training initiatives to help establish a modern and fully capable data science workforce as well as assistance in guiding career paths for data science researchers. DSH affiliates include principally doctorate-level researchers experienced in a wide range of areas needed to develop and deliver a robust curriculum for managing data-intensive scientific research. Examples of expertise areas include:

- Data modeling and integration
- Machine learning and graph analytics
- Performance modeling for 'big data' platforms and workloads
- Scalable, high-performance analytics
- Scientific visualization

'HPC Share' – Accelerating the Flow and FAIRness of Data

In March 2020 SDSC announced the launch of 'HPC Share', a data sharing resource that will enable users of the Center's high-performance computing resources to easily transfer, share, and discuss their data within their research teams and beyond. HPC Share is powered by SDSC's open-source SeedMeLab software that was developed with support from the NSF. Its built-in web services, coupled with an API extension, make it a versatile platform to create branded data repositories for small research groups to large communities. Additionally, users can integrate their existing research data flow—serving as a stepping stone for researchers to realize FAIR (findable, accessible, interoperable, and reusable) data management in practice.

"HPC users face a range of hurdles to share their data," said SDSC Visualization Group Leader Amit Chourasia, also SeedMeLab's principal investigator. "Their collaborators may not have adequate context of their data, or may not be able to find or access data from HPC system. This can slow scientific discovery and burden researchers to devise ad-hoc data sharing mechanisms rather than focusing on their research."

HPC Share solves these hurdles by letting users accelerate the pace of research and information exchange via a ready-to-use infrastructure. Its key capabilities include easy data transfer, accessibility and sharing via a web browser on any device, the ability to add annotation to any file or folder as well as discuss and visualize tabular data. HPC Share is available to all users of SDSC's *Comet* supercomputer, with a potential expansion to SDSC's *Expanse* supercomputer to enter production in late 2020.

(Please read about CloudBank, another collaboration to support the advancement of data-enabled research, on page 9)

QRGO.PAGE.LINK/QPBF7

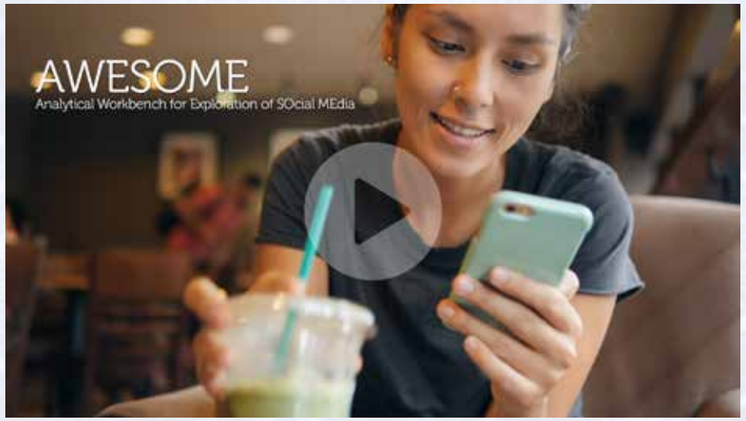
Amarnath Gupta, leader of the AWESOME project, is director of the Advanced Query Processing Lab and a research scientist at SDSC.

'AWESOME' Social Media Data Platform

SDSC researchers have developed an integrative data analytics platform that harnesses the latest 'big data' technologies to collect, analyze, and understand social media activity, along with current events data and domain knowledge. Called AWESOME (Analytical Workbench for Exploring SOcial MEdia), the platform can continuously ingest multiple sources of real-time social media data and scalable analysis of such data for applications in social science, digital epidemiology, and internet behavior analysis. AWESOME is assisting social science researchers, global health professionals, and government analysts by using real-time, multi-lingual, citizen-level social media data and automatically crosslinking it to relevant knowledge to better understand the impact on and reaction to significant social issues.

Two new projects are now being developed on the AWESOME platform. The National Institutes of Health (NIH)-funded TemPredict project seeks to develop machine learning models to predict the onset of COVID-19 in an individual by integrating multiple physiological data from wearable sensors with health surveys. A proof-of-concept project was completed for the U.S. Navy, where a knowledge graph was constructed on the AWESOME platform to discover technology gaps in current research conducted or sponsored by the Navy. A recommendation service was also developed to identify commercial organizations that may bridge the gaps.

Funded by the NSF and NIH, AWESOME's goal is to benefit society through areas as diverse as detecting free speech suppression, to shaping policy decisions or even slowing the spread of viruses. "We are now seeing how new areas of socially aware scientific missions are starting to use and benefit from this platform for societal good," said SDSC researcher Amarnath Gupta, who leads the AWESOME project.



Learn more about the AWESOME social media data platform by visiting SDSC's YouTube channel. Scan the QR code or visit youtu.be/dQk4Hkr5rJY

Subhashini Sivagnanam, principal investigator for Open Science Chain.

Open Science Chain

The National Science Foundation (NSF)-funded Open Science Chain (OSC) is a cyberinfrastructure platform that enables a broad set of researchers to efficiently share, verify, and validate the integrity of published data while preserving the provenance. "The goal of OSC is to increase the confidence of scientific results and enhance data sharing, which in turn leads to greater research productivity and reproducibility," said Subhashini Sivagnanam, principal investigator for the grant and a principal scientific computing specialist with SDSC's Data-Enabled Scientific Computing division. The OSC comprises a consortium blockchain platform that securely stores information about scientific data including its provenance and ownership information, as well as a web portal that lets researchers share, search, and verify the authenticity of datasets. As datasets change or evolve over time, this new information is appended to the OSC blockchain, enabling researchers to view a detailed history of that dataset. On the OSC portal, researchers can also build research workflows linking multiple datasets and code used in their published results, including those maintained in external repositories such as GitHub.

OPENSCIENCECHAIN.ORG

Amit Chourasia, principal investigator for SeedMeLab and leader of the SDSC Visualization Group.

SeedMeLab

SeedMeLab is a cloud service for research teams struggling with intractable data organization and access that disrupts productivity, resulting in a deluge of emails and attachments that obscure discovery and perpetuates poor knowledge retention. Unlike other file sharing services, SeedMeLab provides an effective data organization and sharing platform that empowers collaboration with an ability to add data/research context, discussion, and visualization to any file or folder. Site personalization, branding, and customization establishes distinction to the data while retaining full ownership.

SeedMeLab is available as a managed service from SDSC and the core software is available under an open source license. Customization support and licensing of additional software components are also available. SeedMeLab use cases include collaboration hubs, data management plan implementations for grants, data repositories, and science gateways.

SEEDMELAB.ORG

SDSC CENTERS OF EXCELLENCE

SDSC's Centers of Excellence are part of a broad initiative to assist researchers across many data-intensive science domains, including those that are relatively new to computational and data-enabled science. These centers represent key elements of SDSC's wide range of expertise, from big data management to the analysis and advancement of the internet.

SDSC Chief Data Science Officer Ilkay Altintas also directs the WorDS Center and is principal investigator for the WIFIRE Lab.

WORKFLOWS FOR DATA SCIENCE

Called WorDS for 'Workflows for Data Science', this center of excellence combines over a decade of experience within SDSC's Scientific Workflow Automation Technologies Laboratory, which developed and validated scientific workflows for researchers working in computational science, data science, and engineering. "Our goal with WorDS is to help researchers create their own workflows to better manage the tremendous amount of data being generated in so many scientific disciplines, while letting them focus on their specific areas of research instead of having to solve workflow issues and other computational challenges as their data analysis progresses from task to task," said SDSC Chief Data Science Officer Ilkay Altintas, also director of WorDS. Funded by a combination of sponsored agreements and recharge services. WorDS' expertise and services include:

- World-class researchers and developers with expertise in data science, big data, and scientific computing technologies
- Research on workflow management technologies that led to the collaborative development of the popular Kepler Scientific Workflow System
- Development of data science workflow applications through a combination of tools, technologies, and best practices
- Hands-on consulting on workflow technologies for big data and cloud systems, i.e., MapReduce, Hadoop, Yarn, Spark, and Flink
- Technology briefings and classes covering end-to-end support for data science

WORDS.SDSC.EDU

Sandeep Chandra is director of SDSC's Sherlock Division.

SHERLOCK

SDSC's Sherlock Division provides HIPAA, FISMA, and CUI-compliant cloud and data cyberinfrastructure to meet the secure computing and data management requirements throughout academia, government, and industry. In mid-2020 the division expanded its multi-cloud solution, Sherlock Cloud, to include the Google Cloud Platform. The expansion secures the trifecta of major public commercial cloud platforms included within Sherlock's solution offerings – Sherlock also partners with Amazon Web Services and Microsoft Azure, and offers an on-premise secure hosting capability at UC San Diego. The division also broadened its secure cloud solutions portfolio with the launch of Skylab, an innovative customer-owned cloud platform solution that provides a self-standing, compliant environment for secure workloads in the Amazon Web Services Cloud. Skylab leverages Sherlock's expertise in NIST 800-53 and HIPAA approved templates. "Our research showed that many organizations had several constraints when building cloud environments or buying managed services to host protected data, including sense of ownership, lack of knowledge regarding regulatory data management, and budgetary concerns," said Sandeep Chandra. "Skylab delivers a low-cost, cloud-based, solution that allows customers' IT departments to own, build, and manage their own secure computing enclaves in the cloud."

SHERLOCK.SDSC.EDU

KC Claffy is director of CAIDA, and a resident research scientist at SDSC.

CAIDA

The Center for Applied Internet Data Analysis (CAIDA) was the first center of excellence at SDSC. Formed in 1997, CAIDA conducts network research and builds research infrastructure to support large-scale data collection, curation, and data distribution to the scientific research community. CAIDA's founder and director, KC Claffy, is a resident research scientist at SDSC whose research interests span internet topology, routing, security, economics, future internet architectures, and policy. CAIDA was recently awarded a $4 million, five-year NSF grant to integrate several of its existing research infrastructure measurement and analysis components into a new Platform for Applied Network Data Analysis, or PANDA. The platform, to include a science gateway component, is in response to feedback from the research community that current modes of collaboration do not scale to the burgeoning interest in the scientific study of the internet.

In late 2019 CAIDA received a $1 million Convergence Accelerator planning grant from the NSF to evaluate the feasibility of codifying an Open Knowledge Network about properties of the internet identifier system – the domain names and addresses that represent communication entities – and the structural relationships among these entities. The ultimate goal is to address long-standing gaps in consumer protection and cybersecurity operations and research. "Despite herculean efforts across industry, government, NGOs, and academia, we still lack an understanding of the effectiveness of risk-mitigating efforts, or to what extent such defenses have been deployed," said Claffy. The CAIDA project, called OKN-KISMET for Open Knowledge Network - Knowledge of Internet Structure: Measurement, Epistemology, and Technology, focuses on two key tasks. The first is a team-building effort led by initial partners with a strong history of navigating the interdisciplinary challenges of internet mapping research, including commercial and privacy sensitivities, notably evidence of vulnerabilities or harm to businesses, consumers, and the infrastructure itself. The second leverages a set of use cases prioritized by the team to undertake the design and prototyping necessary to explore the technical feasibility of the proposed Open Knowledge Network.

CAIDA.ORG

James Short is director of BlockLAB and SDSC lead scientist.

CENTER FOR LARGE SCALE DATA SYSTEMS

In late 2018 CLDS formally opened a new blockchain research laboratory with the objectives of exploring the principal technologies and business use cases in blockchains, distributed ledgers, digital transactions, and smart contracts. Founded in partnership with technology firms AEEC, Collibra, Decision Sciences, Dell Technologies, IBM, Intel, and Luna PBC, the new laboratory, called BlockLAB, is conducting research in blockchain and distributed ledger technologies (DLT) and their potential business applications across a wide range of industrial and organizational settings. "One of our primary goals is to work closely with industry partners to provide foundational knowledge to help science-based and industrial companies evaluate the potential benefits and risks of applying these new technologies to critical, large-scale transaction and data-intensive business processes," said James Short, SDSC lead scientist and BlockLAB's director.

CLDS.INFO/BLOCKLAB.HTML

FOCUSED SOLUTIONS& APPLICATIONS

FOR **DATA-DRIVEN DISASTER RELIEF**

Earthquakes, wildfires, and severe rainstorms are unfortunately all-too-common occurrences in California as well as in other parts of the world. SDSC researchers have teamed up with scientists at UC San Diego and beyond to work with first-responder agencies and government officials to support and enhance a range of public safety programs using a key resource: collecting real-time data and using predictive analytics and other tools to help mitigate the often tragic effects of such events.

Gathering and analyzing such data is possible via internet-connected cyberinfrastructures to detect both ground and weather conditions that could potentially harm people and property in populated and more remote areas. In short, data equals knowledge, and SDSC is at the forefront with a range of advanced technologies to support public safety.

WIFIRE.UCSD.EDU

WIFIRE Lab Expands Partnerships, Receives NSF Artificial Intelligence Grant

Started under a multi-year, $2.65 million National Science Foundation (NSF) grant, SDSC's WIFIRE cyberinfrastructure attracted high levels of interest among first-responders during the latest fiscal year as a resource to help them better monitor, predict, and mitigate wildfires by analyzing forecasts from the National Weather Service, vegetation readouts from the U.S. Department of Interior, and satellite data from NASA. WIFIRE Lab is now an all-hazards data resource, enabling analysis and modeling to turn data into knowledge across environmental hazards, as most of the time hazards are connected, such as wildfires and debris flows.

The primary mission of the WIFIRE Lab is to turn data and computing into a utility for advancing fire science and its application to practice. WIFIRE is designed to use any fire modeling software in combination with data at scale. It also provides open data catalogs that integrate and serve data from diverse providers, some of which fund WIFIRE to sustain this data catalog. Through an automated process, the collected data is prepared to enable more accurate modeling of fire and other environmental hazards.

The NSF recently awarded WIFIRE researchers at SDSC and three other universities around the country a grant valued at more than $900,000 to develop WIFIRE Commons, a data-driven initiative intended to leverage advanced artificial intelligence (AI) techniques to limit or even prevent the devastating effects of wildfires, many of which have been ravaging numerous parts of the U.S. well ahead of the typical wildfire season. Partner organizations under the new NSF award, part of the agency's NSF's Convergence Accelerator Phase I project, include the University of Southern California, Los Alamos National Laboratory, and Tall Timbers Research Station.

"When it comes to fire modeling, one size does not fit all," said SDSC Chief Data Scientist Ilkay Altintas, principal investigator for WIFIRE. "This new award will advance WIFIRE's ability and accuracy by allowing us to explore and add the latest AI techniques to our extensive use of predictive data analytics and overall cyberinfrastructure. Looking ahead, our vision is that WIFIRE CI and AI will be useful in developing technologies to understand and combat other disasters, such as mapping the spread of floods, the spread of smoke plumes in fires, and other areas where disaster mitigation is of the highest priority." *(Read more about WIFIRE's activity to help combat California wildfires on page 17)*

Credit: WIFIRE Lab

SDSC Distinguished Research Scientist Hans-Werner Braun is a co-founder of HPWREN.

First Responder Firefighting Efforts Continue Expansion beyond San Diego Area

During the latter half of 2019 and into 2020, numerous upgrades were made to the High Performance Wireless Research and Education Network (HPWREN) network as well as the more recent ALERTWildfire network. ALERTWildfire currently has more than 600 cameras distributed throughout California. Fifty-three ALERTWildfire cameras are using HPWREN infrastructure at 35 locations in San Diego, Orange, Los Angeles, Santa Barbara, and Riverside counties.

Co-founded in 2000 by Hans-Werner Braun, a research scientist at SDSC, and Frank Vernon, a research geophysicist with Scripps Institution of Oceanography, and initially funded by the NSF, HPWREN has since transitioned to a user-funded system and has undergone a significant expansion and upgrading to cover hundreds of miles throughout southern California's most rugged terrain. Today, the HPWREN cyberinfrastructure connects even more hard-to-reach areas in remote environments and includes a system of internet-based cameras and weather stations to report local weather and environmental conditions including storms, wildfires, and earthquakes, to local authorities.

"Because we were able to determine post-fire that the smoke produced in the incipient phases of the Holy Fire was visible on these cameras prior to the initial 911 call, we plan on using this technology as 'virtual fire towers' on identified 'high-hazard' days to increase situational awareness," said Brian Norton, Division Chief – Special Operations, Orange County Fire Authority, during an HPWREN Users' Group meeting in mid-2019. "This will be done in partnership with volunteer fire watch patrols already established in Orange County."

HPWREN.UCSD.EDU

Following repair work, Jim Hale (front) and Adam Brust align an HPWREN antenna between Mount Laguna and the Palomar Observatory. Credit: HPWREN.

Senior Computational Scientist Yifeng Cui directs SDSC's High-Performance Geocomputing Laboratory.

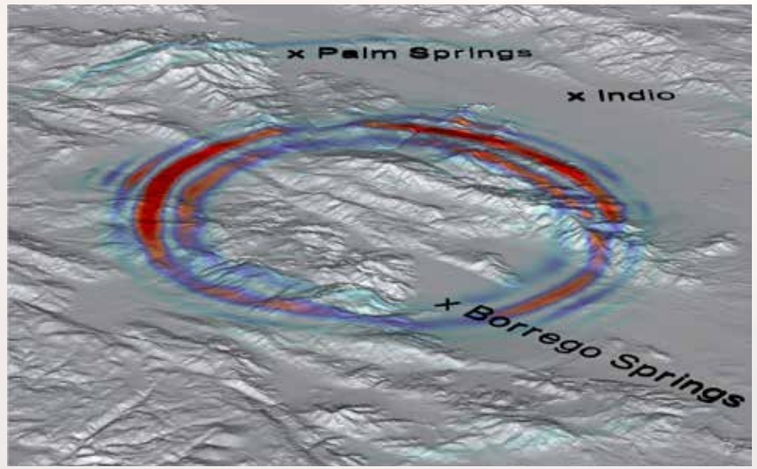
Seismic Shifts Using Supercomputing

In 2019 researchers at SDSC and Intel developed an innovative software package that gave seismic modeling capabilities a notable advancement. Called Extreme-Scale Discontinuous Galerkin Environment, or EDGE, the software is designed to take advantage of the latest generation of Intel processors. The project is the result of a joint initiative that also included the Southern California Earthquake Center (SCEC), one of the largest open research collaborations in geoscience. The findings were presented in June 2019 at the annual International Supercomputing (ISC) High-Performance Conference in Frankfurt, Germany.

Some highlights of EDGE include:

- Speed: EDGE is among the fastest seismic simulations yet devised, capable of 10.4 PFLOPS (Peta Floating-point Operations Per Second, or one quadrillion calculations per second). This allows more simulations at the same frequency so researchers can explore scenarios in much more detail.
- Scalability: EDGE's compute infrastructure runs at a large scale, letting researchers increase the modeling frequency range.
- Simultaneous Events Rendering: EDGE can simulate multiple events in one software execution, providing the ability to apply similarities in setups for efficiency, such as sharing mountain topography in multiple simulations. Researchers can run two to five times more simulations to make the most of their resources.

"Cloud services provide a combination of infrastructure flexibility and scalability that EDGE requires for its groundbreaking modeling, which helps to gain the insight needed for better preparation before a major earthquake occurs, and assists disaster recovery and relief efforts afterward," said Yifeng Cui, founding director of the High-Performance GeoComputing Laboratory at SDSC, and principal investigator for SDSC and SCEC.



ORGO.PAGE.LINK/TWSMR

Example of hypothetical seismic wave propagation with mountain topography using the EDGE software. Shown is the surface of the computational domain covering the San Jacinto fault zone between Anza and Borrego Springs in California. Colors denote the amplitude of the particle velocity, where warmer colors correspond to higher amplitudes. Credit: Alex Breuer, SDSC.

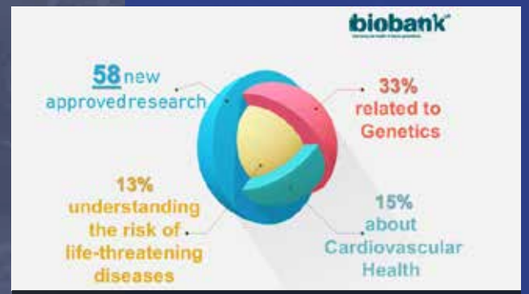
SDSC INDUSTRY RELATIONS

Ron Hawkins is director of Industry Relations for SDSC and manages the Industry Partners Program, which provides member companies with a framework for interacting with SDSC researchers and staff to develop collaborations.

BUILDING CAPABILITY TO STRENGTHEN INDUSTRIAL PARTNERSHIPS

For the SDSC Industrial Partners Program, the period 2019-2020 was one of new challenges but one that also provided capability-building opportunities for the future. At its core, developing university-industry partnerships is a business of relationship-building and as social creatures, humans are accustomed to doing this by meeting, shaking hands, and discussing face-to-face. The emergence of a global viral pandemic interrupted the usual process of engagement and required rethinking how to manage the usual activities while working from home using online collaboration tools. As we learned how to work in this new environment and wondered when we might be able to return to the 'old normal,' we continued to service existing partnerships as well as adding major capabilities forming a foundation for an exciting round of new partnerships to be developed in 2021 and beyond.

For a large pharmaceutical company, we implemented a major storage expansion using our Universal Scale Storage (USS) service, founded on technology from Qumulo. This storage expansion will permit corporate researchers to conduct genomics-related studies utilizing data from population-scale genome sequencing projects such as UK Biobank and other sources. SDSC experts will assist in the development of data integration strategies, analysis pipelines, and HPC-Cloud integration to support these studies. For a division of a global industrial conglomerate, SDSC is providing high-performance computing capacity on its *Comet* and *TSCC* supercomputers to support research into new industrial materials. At the other end of the spectrum in terms of business size, SDSC continued to collaborate with local startup companies such as Arima Genomics and GigaIO, providing HPC services for genome analysis and collaborating on development of composable computing systems based on a new networking fabric as well as GPU and FPGA computational accelerators.



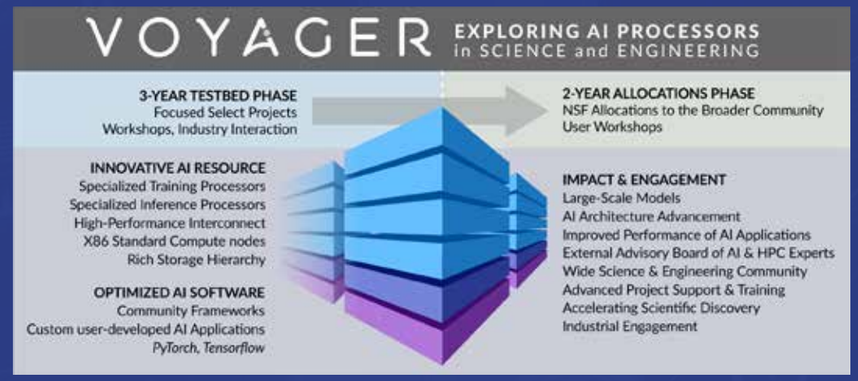
UK Biobank recruited 500,000 people in the 40-69 age range in 2006-2010 from across the UK to provide detailed health information about themselves and have their health tracked.

The period 2019-2020 saw the award of new supercomputer capabilities that will benefit industrial partnerships as well as academic research. SDSC is taking action to augment its new *Expanse* supercomputer with additional capacity to support industrial collaborations. This augmentation will provide state-of-art CPU and GPU processing capabilities to conduct modeling and simulation, data analytics, and AI/machine learning (ML)-based techniques for both industry-sponsored and industry-conducted projects. As usual, SDSC experts will be on tap to help with application configuration, code tuning/optimization, workflow development, domain expertise, and other activities.

Also announced during the 2019-2020 period was the award of the *Voyager* supercomputer, which will be focused on providing high-performance computation for AI and ML. Deployment of *Voyager* will begin in mid-2021. In advance of the *Voyager* award, SDSC established the AI Technology Lab (AITL) in October 2019, which is led by Ron Hawkins and will provide a framework for leveraging *Voyager* to foster new industry collaborations aimed at exploring emerging AI and ML technology for scientific and industrial uses, while also helping to prepare the next-generation workforce. Outreach activities for developing new industrial partnerships under AITL and *Voyager* are planned to commence in calendar year 2021.

In summary, the period 2019-2020 was one of learning to conduct business in new ways, maintaining a steady course with strong partners, and building a strong foundation for the future of the program.

INDUSTRY.SDSC.EDU



Overview of *Voyager*, a high-performance, innovative resource to be developed by SDSC for conducting AI research across a wide range of science and engineering domains.

BY THE NUMBERS
FY2019/20

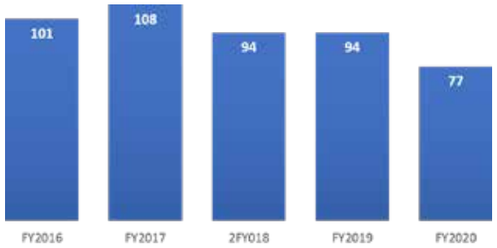
Proposal Success Rate

	FY16	FY17	FY18	FY19	FY20
Proposals Submitted	73	84	75	67	71
Proposals Funded	33	33	26	31	26
Success Rate	45%	39%	35%	46%	37%

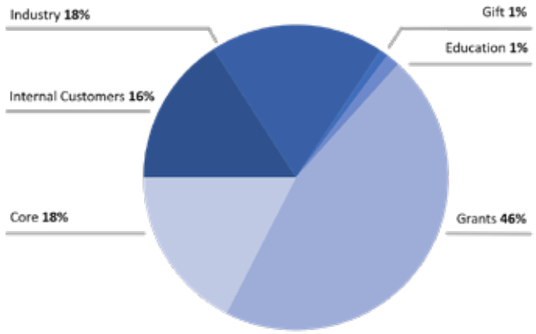
In perhaps the most competitive landscape for federal funding in the last two decades, SDSC's overall success rate on federal proposals averages 40 percent over the last five years compared to the FY2020 national average of about 23 percent for Computer and Information Science and Engineering proposals at the National Science Foundation.

During its 36-year history, SDSC revenues have well exceeded $1 billion, a level of sustained funding matched by few academic research units in the country. As of the start of FY2020, SDSC had 90 active awards totaling about $126 million.

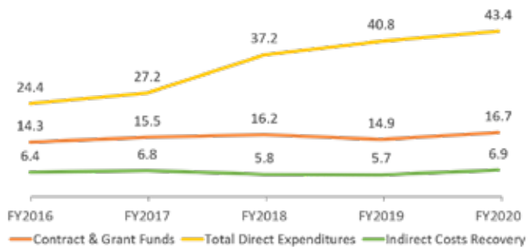
Number of Sponsored Research Awards



Revenue Categories: $36.5 Million



Total Expenditures ($ Millions)



LEADERSHIP
EXECUTIVE TEAM

Michael Norman
SDSC Director

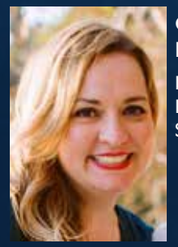
Shawn Strande
Deputy Director

Ilkay Altintas
Chief Data Science Officer

Chaitanya Baru
(on assignment to NSF)
Associate Director, Data Science and Engineering

Sandeep Chandra
Division Director, Sherlock

Ronald Hawkins
Director, Industry Relations



Christine Kirkpatrick
Division Director, Research Data Services

Samuel "Fritz" Leader
Chief Administrative Officer

Amit Majumdar
Division Director, Data-Enabled Scientific Computing

Frank Würthwein
Lead, Distributed High-Throughput Computing

Michael Zentner
Division Director, Sustainable Scientific Software

Jan Zverina
Division Director, External Relations

RESEARCH EXPERTS
SDSC Computational & Data Scientists

Ilkay Altintas, Ph.D
Chief Data Science Officer, SDSC
Director, Workflows for Data Science (WorDS) Center of Excellence
Lecturer, Computer Science and Engineering, UC San Diego
Scientific workflows
Big data applications
Distributed computing
Reproducible science

Chaitan Baru, Ph.D.
SDSC Distinguished Scientist
Director, Center for Large-scale Data Systems Research (CLDS), SDSC
Associate Director, Data Science and Engineering, SDSC
Associate Director, Data Initiatives, SDSC
Data management and analytics
Large-scale data systems
Parallel database systems

Hans-Werner Braun, Ph.D.
Research Scientist Emeritus, SDSC
Internet infrastructure, measurement/analysis tools
Wireless & sensor networks
Internet pioneer (PI, NSFNET backbone project)
Multi-disciplinary and multi-intitutional collaborations

Sandeep Chandra
Executive Director, Sherlock Cloud
Director, Health Cyberinfrastructure Division, SDSC
Compliance (NIST, FISMA, HIPAA), scientific data management
Cloud computing
Systems architecture and infrastructure management

Dong Ju Choi, Ph.D.
Senior Computational Scientist, SDSC
HPC software, programming, optimization
Visualization
Database and web programming
Finite element analysis

Amit Chourasia
Senior Visualization Scientist, SDSC
Lead, Visualization Group
Principal Investigator, SeedMeLab
Visualization, computer graphics
Data management and sharing

KC Claffy, Ph.D.
Director/PI, CAIDA (Center for Applied Internet Data Analysis), SDSC
Adjunct Professor, Computer Science and Engineering, UC San Diego
Internet data collection, analysis, visualization
Internet infrastructure development of tools and analysis
Methodologies for scalable global internet

Melissa Cragin, Ph.D.
Chief Strategist for Data Initiatives, Research Data Services (RDS), SDSC
Shared data infrastructure
Research policy
Scholarly communication

Daniel Crawl, Ph.D.
Associate Director, Workflows for Data Science (WorDS) Center of Excellence
Technical Lead, WIFIRE Lab
Spatial and temporal data integration/analysis, interdisciplinary applications

Yifeng Cui, Ph.D.
Director, Intel Parallel Computing Center, SDSC
Director, High-performance GeoComputing Laboratory, SDSC
Principal Investigator, Southern California Earthquake Center
Senior Computational Scientist, SDSC
Adjunct Professor, San Diego State University
Earthquake simulations, multimedia design and visualization
Parallelization, optimization, and performance evaluation for HPC

Alberto Dainotti, Ph.D.
Assistant Research Scientist, CAIDA (Center for Applied Internet Data Analysis)
Internet measurements, traffic analysis, network security
Large-scale internet events

Diego Davila
Research Scientist, Distributed High-Throughput Computing, SDSC
High-Throughput Computing, globally distributed compute and data federations

Jose M. Duarte, Ph.D.
Scientific Team Lead, RCSB Protein Data Bank
Bioinformatics, structural biology, computational biology

Andreas Goetz, Ph.D.
Co-Director, CUDA Teaching Center
Co-Principal Investigator, Intel Parallel Computing Center
Quantum chemistry, molecular dynamics
ADF and AMBER development
GPU accelerated computing

Madhusudan Gujral, Ph.D.
Bioinformatics & Genomics Lead, SDSC
Processing and analysis of genomics data
Biomarkers discovery
Genomics software tools development
Data management

Amarnath Gupta, Ph.D.
Associate Director, Academic Personnel, SDSC
Director of the Advanced Query Processing Lab, SDSC
Co-Principal Investigator, Neuroscience Information Framework (NIF) Project, Calit2
Bioinformatics
Scientific data modeling
Spatiotemporal data management
Information integration and multimedia databases

Edgar Fajardo Hernando
Research Scientist, Distributed High-Throughput Computing, SDSC
High-Throughput Computing
Globally distributed compute and data federations

Martin Kandes, Ph.D.
Computational and Data Science Research Specialist, SDSC
Bayesian statistics, combinatorial optimization
Nonlinear dynamical systems, numerical partial differential equations

Amit Majumdar, Ph.D.
Division Director, Data-Enabled Scientific Computing, SDSC
Associate Professor, Dept. of Radiation Medicine and Applied Sciences, UC San Diego
Parallel algorithm development
Parallel/scientific code, parallel I/O analysis and optimization
Science gateways
Computational and data cyberinfrastructure software/services

Mark Miller, Ph.D.
Principal Investigator, Biology, SDSC
Principal Investigator, CIPRES Gateway, SDSC & XSEDE
Principal Investigator, Research, Education and Development Group, SDSC
Structural biology/crystallography
Bioinformatics, next-generation tools for biology

Dmitry Mishin, Ph.D.
Research Programmer Analyst, Data-Enabled Scientific Computing, SDSC
Research Programmer Analyst, Calit2
HPC systems, virtual and cloud computing
Kubernetes and containers
Data storage, access, and visualization

Dave Nadeau, Ph.D.
Senior Visualization Researcher, SDSC
Data mining, high-dimensionality data sets, visualization techniques
User interface design, software development
Audio synthesis

Viswanath Nandigam
Associate Director, Advanced Cyberinfrastructure Development Lab, SDSC
Data distribution platforms
Scientific data management
Science gateways
Distributed ledger technologies

Mai H. Nguyen, Ph.D.
Lead for Data Analytics, SDSC
Machine learning
Big data analytics
Interdisciplinary applications

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

Francesco Paesani, Ph.D.
Lead, Laboratory for Theoretical and Computational Chemistry, UC San Diego
Theoretical chemistry, computational chemistry, physical chemistry

Dmitri Pekurovsky, Ph.D.
Member, Scientific Computing Applications group, SDSC
Optimization of software for scientific applications
Performance evaluation of software for scientific applications
Parallel 3-D fast Fourier transforms
Elementary particle physics (lattice gauge theory)

Wayne Pfeiffer, Ph.D.
Distinguished Scientist, SDSC
Supercomputer performance analysis
Novel computer architectures
Bioinformatics

Peter Rose, Ph.D.
Director, Structural Bioinformatics Laboratory, SDSC
Lead, Bioinformatics and Biomedical Applications, Data Science Hub, SDSC
Structural biology/bioinformatics
Big data applications
Data mining and machine learning
Reproducible science

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

Igor Sfiligoi
Senior Research Scientist, Distributed High-Throughput Computing, SDSC
Lead Scientific Software Developer and Researcher, SDSC, UC San Diego
High-Throughput Computing
Globally distributed compute and data federations
Cybersecurity in globally distributed systems

James Short, Ph.D.
Lead Scientist, SDSC
Co-Director, Center for Large-scale Data Systems Research (CLDS), SDSC
Director, BlockLAB, SDSC
Market adoption, business application and economics of blockchain
Valuation of data and information markets
Textual analysis of data (Natural Language Processing and machine learning)
Corporate data policy and regulatory/legal frameworks

Robert Sinkovits, Ph.D.
Director, Scientific Computing Applications, SDSC
Director, SDSC Training
Co-Director for Extended Collaborative Support, XSEDE
High-performance computing,
Software optimization and parallelization
Structural biology, bioinformatics
Immunology

Subhashini Sivagnanam
Senior Computational and Data Science Specialist,
Data-Enabled Scientific Computing Division, SDSC
HPC solutions and applications
Science gateways
Data reproducibility
Computational and data cyberinfrastructure software/services

Shava Smallen
Manager, Cloud and Cluster Development
Principal Investigator, Pacific Rim Application and Grid Middleware Assembly (PRAGMA)
Deputy Manager, XSEDE Requirements Analysis and Capability Delivery (RACD) group
Cyberinfrastructure monitoring and testing
Cloud infrastructure tools
Cluster development tools

Mahidhar Tatineni, Ph.D.
User Support Group Lead, SDSC
Research Programmer Analyst
Optimization and parallelization for HPC systems
Aerospace engineering

Mary Thomas, Ph.D.
Computational Data Scientist, SDSC
CyberTraining
Cyberinfrastructure and emerging technologies
HPC software and portal development

Igor Tsigelny, Ph.D.
Research Scientist, SDSC
Research Scientist, Department of Neurosciences, UC San Diego
Computational drug design
Personalized cancer medicine
Gene networks analysis
Molecular modeling/molecular dynamics
Neuroscience

David Valentine, Ph.D.
Research Programmer, Spatial Information Systems Laboratory, SDSC
Spatial and temporal data integration/analysis
Geographic information systems
Spatial management infrastructure
Hydrology

Frank Würthwein, Ph.D.
Lead, Distributed High-Throughput Computing, SDSC
Executive Director, Open Science Grid
Professor of Physics, UC San Diego
High-capacity data cyberinfrastructure
High-energy particle physics

Ilya Zaslavsky, Ph.D.
Director, Spatial Information Systems Laboratory, SDSC
Spatial and temporal data integration/analysis
Geographic information systems, geosciences
Visual analytics

Michael Zentner, Ph.D.
Director, Sustainable Scientific Software, SDSC
Director, HUBzero Platform for Science and Engineering
Co-Principal Investigator, Network for Computational Nanotechnology
Principal Investigator, Science Gateways Community Institute (SGCI)
Sustainable scientific software
Chemical engineering

Andrea Zonca, Ph.D.
HPC Applications Specialist
Data-intensive computing
Computational astrophysics
Distributed computing with Python
Jupyterhub deployment at scale

SDSC @ UC San Diego

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

www.sdsc.edu
email/info@sdsc.edu
twitter/SDSC_UCSD
instagram.com/SDSC_UCSD
facebook/SanDiegoSupercomputerCenter
youtube.com/SanDiegoSupercomputerCenter