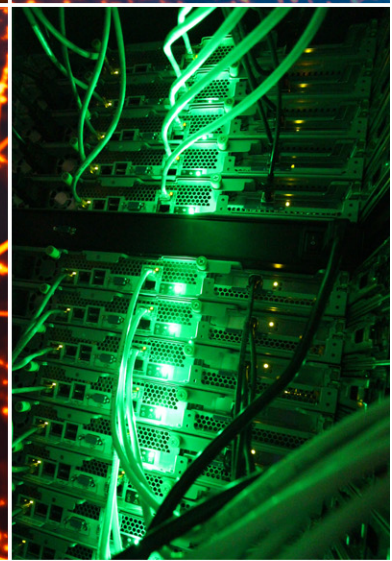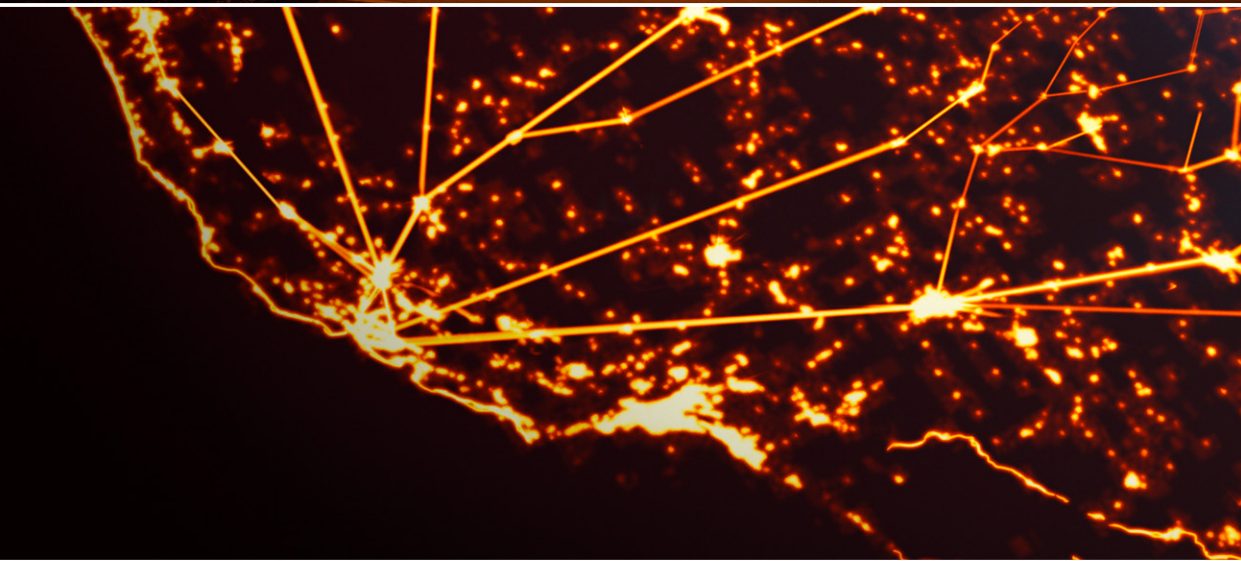# UC@SDSC Strategic Overview
2014 - 2016

SAN DIEGO SUPERCOMPUTER CENTER at UC SAN DIEGO

Executive Summary

Today, science is sitting at a crossroads between two dynamic and transformative eras in the history of research: sometimes referred to as the "third and fourth paradigms" of science.

For almost three decades, the San Diego Supercomputer Center (SDSC) helped blaze a trail in the "third paradigm," built around advances in computational science that were directed to solve the grand challenges of its day. The design and construction of bigger and faster computers capable of millions, then billions, trillions and now quadrillions of calculations per second are enabling these third paradigm efforts, following centuries of research built around two earlier trusted fundamental methodologies: experimental science and theoretical science.

Recently, we've pushed into yet another research era called the "fourth paradigm," built around the need to harness what many call "big data" or the "data deluge": the product of ultrafast supercomputers; observational tools such as satellites, oceanographic sensors, astronomical observatories; and personal websites. The planet is awash in data, so much so that it threatens our abilities to preserve, analyze, comprehend, and apply in a meaningful way. "Data-enabled science and engineering" has emerged as academia's response to this fourth paradigm conundrum.

The National Science Foundation (NSF) recognizes that the two methodologies of modern science – the third and fourth paradigms -- do not exist in isolation but complement one another, and are often practiced in the same labs. NSF has dubbed the merging of this scientific research duo as "cyberinfrastructure computation and data-enabled science and engineering."

SDSC is now bridging these two contemporary paradigms of research in the 21st Century with a new initiative called UC@SDSC, with a goal of leveraging the best that SDSC and UC can offer to help solve the grand challenges of our times.

UC@SDSC provides specific programs that apply the Center's resources, services and expertise in computation and data-enabled science and engineering, built around three fundamental tenets:

COLLABORATION

Since opening its doors in 1985, SDSC has conceived, nurtured, and raised multiple partnerships and collaborations with individuals, communities, and institutions, across a wide spectrum of disciplines and fields of study at UC and beyond. The resources, services and expertise provided by SDSC have been a tremendous magnet for computational and data-oriented scientists, with the Center now serving as a leader/co-leader/participant on research projects with more than 200 UC principal investigators.

INNOVATION

Combined with its long pioneering history in computational science, SDSC is also able to provide a crucial gateway for innovation in the fourth paradigm, built around the need to harness "big data" and the "data deluge" for data-enabled science and engineering. SDSC has committed itself to this cause and is focusing on developing technologies and infrastructure to help researchers make sense of all this data, by leveraging SDSC's deep expertise in large-scale systems, data management and data integration, data science workflows, data mining and analysis, robust software, and more.

EDUCATION

A vital part of SDSC's guiding philosophy is the empowerment of science and engineering communities in the present, and the development of scientists, technologists and engineers in the next generation. SDSC offers hands-on training in high-performance computing and data science to UC students and researchers in the use of high-performance computing, data-intensive computing, and data analytics focusing on a variety of research disciplines, from neuroscience and geophysics to the humanities, arts and social sciences.

In short, SDSC is identifying and embracing new opportunities created by the emergence of cyberinfrastructure computational and data-enabled science and engineering, and is seeking collaborations and partnerships among its UC counterparts to enhance this effort for the university system-wide and the residents of California. This document provides an overview of a three-year effort to implement this strategy.

Introduction

Our nation's research universities, a traditional source of innovation and discovery, are being challenged by the business community and government leaders to develop new tools to better mine, model, manage, and otherwise analyze a tidal wave of data stemming from our academic centers, commercial laboratories, government scientists, and observational tools such as satellites, oceanographic sensors, astronomical observatories, and social networking sites.

These leaders are trying to better understand the concept of big data, initially at least, to separate hype from reality. In the process, they hope this conversation yields better narratives and/or visualizations to help explain events of importance to science and society, from the seeming fickleness of the marketplace to how virulent diseases spread from neighborhood to neighborhood, city to city, and beyond.

The term 'data science and engineering' has emerged as academia's answer to the big data need, (also known as the fourth paradigm of science) much the way 'computational science and engineering' programs sprung up a couple of decades ago in response to an advancement known as supercomputing (also known as the third paradigm of science). NSF refers to the merger of these two contemporary research paradigms as 'cyberinfrastructure computation and data-enabled science and engineering'.

Today, both scientific methodologies are fundamental components of the same tool kit that researchers need to discover new concepts and advance innovative technologies. With an eye toward bringing the latest cyberinfrastructure computation and data-enabled science and engineering to the UC-wide community, SDSC has launched UC@SDSC. The Center's vast array of data and computational science capabilities includes its supercomputing resources, advanced networking, data storage infrastructure, workflow tools and visualization techniques. A critical mass of expertise across numerous specialties support and integrate these various components into a holistic working environment.

Context

Though SDSC was founded by the NSF almost three decades ago in response to academic researchers nationwide seeking the latest supercomputing resources to solve "grand challenge" problems facing science and society, the Center's mission has evolved over time to include substantial work with, and for, the University of California's initiatives and researchers.

As part of that effort, SDSC has worked to align itself with three UC principles that "define the goals and purpose that drive and distinguish UC-wide research investments." Those principles are:

- *Act as one system of multiple campuses to enhance UC's influence and advantage*

- *Promote efficient inter-campus collaborations and system-wide economies of scale*

- *Serve the State of California*
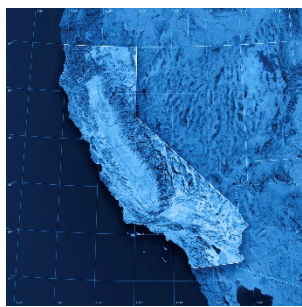
Indeed, when SDSC became an organized research unit at UC San Diego in 1996, then Gov. Pete Wilson recognized the Center as a "cornerstone of California's vision for business and academic leadership in science and technology." Since then, SDSC not only has continued its highly successful national mission, the Center also has delivered value and prestige to UC and the State of California through numerous activities with a focus on its resources, services and expertise in computational and data science. Some notable achievements include:

- **Research and discovery.** During the past five years, SDSC staff has collaborated on projects with more than 90 UC researchers across eight campuses. SDSC staff is currently participating in roughly 50 research grants and proposed collaborations across the UC system.

- **Funds through successful grant applications and research partnerships.** Over its lifespan, SDSC revenues have exceeded $1B—a level of sustained funding matched by few academic research units in the country. SDSC has leveraged the operating funds it receives from the UC Office of the President (UCOP) and UC San Diego and has returned six times the amount in sponsored research awards from FY2009 to FY2013.

- **Data-intensive and High Performance Computing (HPC).** Since its launch in 1985, more than 7,500 UC researchers have used SDSC's HPC systems. The Center currently provides compute and HPC storage resources to about 560 UC researchers, including 150 principal investigators across eight UC campuses.

- **Cost efficient colocation and "green" computing.** More than 90 UC groups spanning eight campuses utilize SDSC's 19,000 square foot data center available as a recharge-based colocation, with an estimated annual system-wide utility savings that exceeds $500K and equivalent to a reduction of 5,500 tons of carbon dioxide.

- **Innovative cyberinfrastructure (CI).** CI services, available at competitive rates to UC researchers, include the Triton computer resource which through 2012 provided more than a million core-hours to more than 600 users across eight UC campuses, and was used for teaching classes both

at UC San Diego and UC Santa Barbara; in addition to the SDSC Cloud storage service, among the largest academic cloud storage systems in the world.

- **Internet measurement, security and data analysis.** Formed in 1997, SDSC's Center for Applied Internet Data Analysis (CAIDA) seeks, through the collection and curation of strategic internet data sets and freely available tools and analysis methodologies, to improve the scientific integrity of network research and to promote more informed engineering, business, and policy decisions regarding the Internet infrastructure.
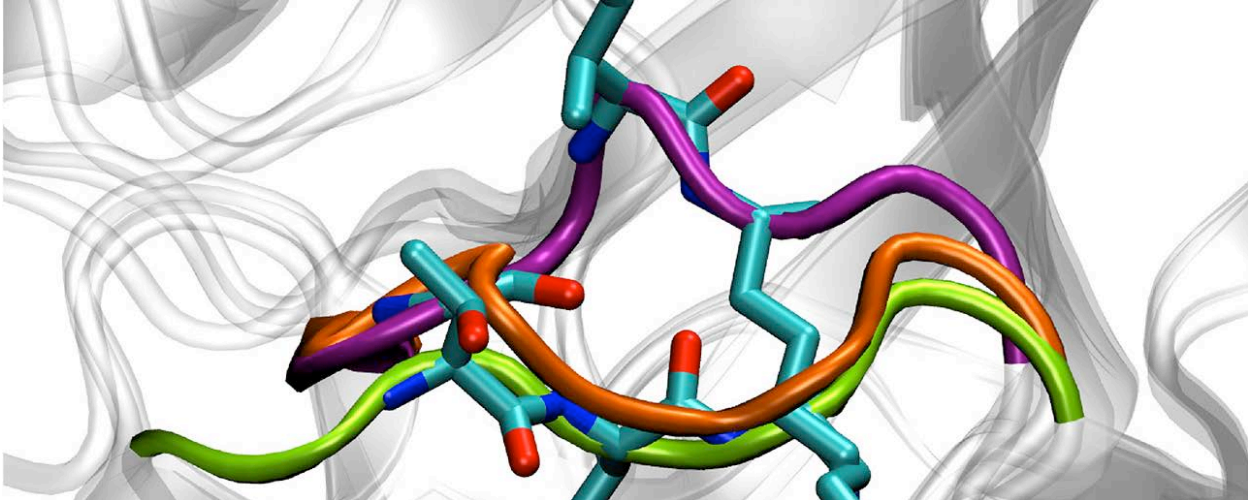
- **Host to numerous critical databases.** SDSC houses numerous data sources of high impact including: the Protein Data Bank (protein structures); Medicaid data from the Centers for Medicare and Medicaid Services (CMS); the Library of Congress Chronopolis digital preservation system; Open Topography.org (LiDAR data); and the American Red Cross Safe and Well website, developed at SDSC in urgent response to the Hurricane Katrina disaster and the need to match missing people with family and friends.

- **Wireless network for science and society.** HPWREN, the High-Performance Wireless Research and Education Network, provides high-speed Internet access to field researchers from disciplines including geophysics, astronomy, and ecology, educational opportunities through connections to learning centers in several communities, and advanced warning and monitoring systems to firefighters in distant sections of San Diego County. HPWREN consists of a collaboration of researchers at SDSC, the Scripps Institution of Oceanography and San Diego State University.

- **Bridge to empower the next generation.** Through its award-winning TeacherTech program, SDSC has trained more than 1,000 teachers in the San Diego region in science and technology, helping many underserved students to span the 'digital divide' to the Information Age. SDSC also provides programs to train and educate local high school students in computer science and technology, while fostering opportunities to bridge the gender gap for women in science, technology, education and math (STEM).

Aim

The document's aim is to set a broad strategic direction for UC@SDSC, and to enunciate a range of goals and activities. A separate document presents specific implementation plans on an annual basis.

Goals and Proposed Activities

GOAL 1

To serve as a foundation of support for research collaborations and partnerships, providing the latest computational infrastructure and data expertise for UC system-wide researchers across a wide spectrum of disciplines.

Activity 1
UC-wide Collaborative Research Opportunity
SDSC experts interested in collaborating with UC researchers can apply for a CRO mini-grant to support collaborative work, leading to an extramural grant proposal.

Activity 2
External Advisory Board (EAB)
These are designated UC faculty/researchers who will proactively seek out collaborations with SDSC and act as ambassadors on their home campuses. The EAB will foster multi-UC collaborations and facilitating support for multi-campus infrastructure, collaborations and partnerships requiring SDSC resources, services and expertise.

Activity 3
'Made in UC' Data Science Cluster
Showcase, extend and benchmark software, tools and applications developed by UC faculty and researchers with benchmarking, extensions and new uses cases documented by SDSC.

GOAL 2

To encourage, promote and support innovative research and technology that advances the curiosity of scientists and California's needs to promote the health and welfare of its citizens.

Activity 1.
Technical expertise
At the core of SDSC's innovation is its critical mass of experts who span numerous specialties in HPC and data science. SDSC experts are ready to partner with UC research collaborators and assist them in utilizing the significant computational, software, and data resources available to address complex research problems
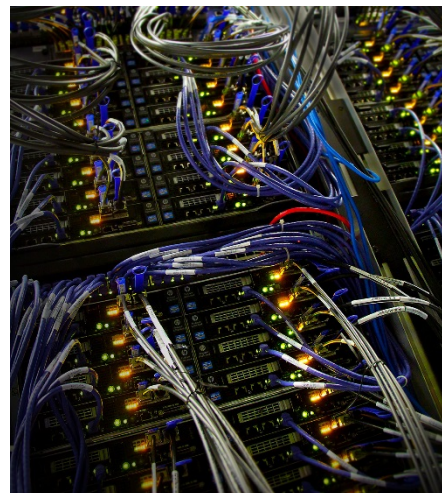

Activity 2
Access to high-performance computing via *Comet*
Comet, a new NSF-funded petascale supercomputer, is

dedicated to the "long tail of science," and expands access and capacity among traditional as well as non-traditional research domains, including "gateways communities." Among its capabilities is support for virtualization of HPC systems to allow users to install and execute custom software stacks. Additional libraries also will be built to extend functionality, modifications made to work in the HPC virtualized Comet environment.

Activity 3
Cyberinfrastructure, as part of an Integrated Digital Infrastructure (IDI)
Half of SDSC's 19,000-square foot data center is available as a recharged-based regional colocation facility for the UC system. The SDSC data center hosts several leading-edge resources including: the Triton Shared Computing Cluster (TSCC), a data, intensive, high-performance computer featuring massive data analysis and preservation for UC researchers; SDSC Cloud, an OpenStack-based computing and object storage cloud, with the latter offering north/south geographical replication, system administration services and data hosting, and cost-effective solutions for long-term data-sharing needs now mandated by federal agencies; Sherlock Cloud, offering FISMA and HIPAA compliant, managed infrastructure and services focusing on a broad range of services, including analytics, case management, secure cloud hosting, and data management; 'Made in UC' Data Science Cluster, a test system at SDSC with a variety of science-related tools and technologies that are readily available for use, including the Hadoop software ecosystem, as well as UCI-developed AsterixDB.

GOAL 3

Empower science and engineering communities in the present, and the development of scientists, technologists and engineers in the next generation, by establishing professional workshops, formal university courses in data science and informatics, along with graduate fellowships in computational science and engineering.

Activity 1
SDSC Training Institutes
Provide UC students and researchers hands-on experience in HPC and data science using some of the nation's most advanced computing platforms. These institutes focus on HPC and data management and analytics for a variety of disciplines across the natural and social sciences, engineering, humanities, medicine, management, and education.
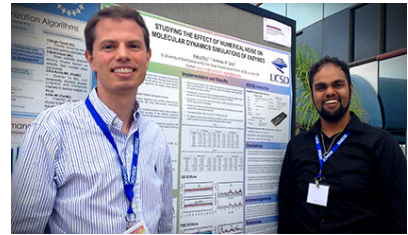
Activity 2
Courses and workshops
To be offered at SDSC, UC San Diego, on-site at other UC locations and online. The short course format will emphasize hands-on training with the latest tools and technologies in HPC and data science.

7

Activity 3

SDSC In-residence Summer Fellowships

8-10 week fellowships will be offered to a select group of UC graduate students, to include room and board and a stipend to work with and be co-mentored by UC and SDSC researchers on specific projects in computational and/or data science.

GOAL 4

Promotes SDSC's commitment to collaborations with UC faculty and researchers, marketing value propositions, identifying opportunities and publishing results.

Activity 1

Create value proposition for UC faculty

Focus on how SDSC partnership increases influence score, award acceptance rate, etc. Work toward a one-stop shopping message, web-based presence for UC PIs; match collaborations that pair software and use cases, suggesting SDSC improvements and extensions of UC-invented software; connect experimental computer scientists with domain scientists (become Pi=shaped); construction of a 'Made in UC' cluster and portal to promote UC software and techniques; benchmarking, use cases, evaluation and inclusion in workflows and scientific pipelines.

Activity 2

Promote activities and achievements

SDSC's external relations team will create newsletters, flyers, and press releases that focus on UC&SDSC activities, services, and achievements, targeting UC researchers as well as the wider user community. Publish success stories on SDSC website.

Activity 3

Conduct public seminars/workshops

Relevant topics on the use of computational and data science will be a focus of public seminars and workshops held at SDSC, other UC campuses and via webinars.

SDSC Governance

SDSC is an organized research unit based at the UC San Diego. The Center's Director reports to the university's Vice Chancellor of Research, and receives advice from an Executive Committee, appointed by the VCR in consultation with SDSC's Director. The Executive Committee meets bimonthly and is charged with actively participating in setting the Center's goals, providing advice on major decisions affecting SDSC, and evaluating critically, the Center's research focus, membership, and outreach programs. As an ORU of UC San Diego, SDSC also is reviewed every five years by an external committee that evaluates its performance and its plans for the next five years.

UC External Advisory Committee

SDSC's UC External Advisory Committee serves as a focal point for counsel and support for multi-campus research collaborations across the UC system-wide, helping to foster and promote research, technological and educational initiatives of critical importance to SDSC, UC system-wide and the residents of California. Members of this committee are charged with: identifying specific government and privately-financed research projects of critical importance to SDSC, UC and the State of California that should be targeted with multi-campus UC collaborations with SDSC; serving as ambassadors to their campus affiliates, fostering collaborations and partnerships with faculty and other researchers on multi-campus research projects of strategic importance to SDSC, UC system-wide and the State of California; facilitating support for multi-campus infrastructure, collaborations and partnerships requiring SDSC resources, services and expertise of strategic value to SDSC, UC system-wide and the State of California.

SDSC

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

www.sdsc.edu