SDSC Distinguished Scientist Wayne Pfeiffer was recently named SDSC’s second “Pi Person of the Year” for his research on computationally challenging problems in bioinformatics, particularly ones in genomics that can benefit from access to data-intensive supercomputers. Named after the π symbol, this award recognizes researchers who have one ‘leg’ in computational science and the other in cyberinfrastructure technology.

Pfeiffer’s latest research includes analyses of genomic variation within species, or even individuals and phylogenetic analyses of evolution among groups of genetically related species. With a background in computational physics and a Ph.D in engineering science from the California Institute of Technology, he began his career at General Atomics, focused on fission and fusion energy. As the longest serving SDSCer, Pfeiffer helped co-found SDSC with Sid Karin in 1985.

[continued on pg. 7]
Welcome to the February/April issue of SDSC Innovators. We’re well into the New Year and we’ve hit the ground running! In keeping with the newsletter’s theme of highlighting the people behind the research, we’ve named our second annual “π Person of the Year”—Wayne Pfeiffer.

Wayne is one of SDSC’s Distinguished Scientists, a bioinformatics researcher who’s been with SDSC from the beginning—he helped write the proposal to found this center almost 30 years ago. Wayne is currently conducting leading-edge research in genome sequencing and has participated in widely publicized projects such as one that studied the genes of a woman who led a healthy life up until the age of 115.

Also on the “people” front, UC San Diego Physics Professor Frank Würthwein has joined SDSC as head of our Distributed High-Throughput Computing Group. A recognized expert in high-energy particle physics, Frank’s mission is to develop and deploy a high-capacity data cyberinfrastructure across the UC system. One key benefit of such a network is that individual PIs at all UC campuses will have direct access to SDSC’s expertise and resources from their home institutions.

In this issue you’ll read about SeedMe, the brainchild of our visualization expert Amit Chourasia. What’s unique here, besides the project itself, is that the article was written by two students who participated in last year’s Research Experiences for High School Students (REHS) program. According to those students, SeedMe may “be just what is needed to weave a larger web of knowledge while drawing the realm of science ever closer.”

Great news on the resources front—Comet has entered early operations! The previous issue of SDSC Innovators has a feature story about our newest supercomputer, and we look forward to hosting its formal rollout later this year. And don’t forget to check out SDSC’s brand new website! More details on this project can be found inside this latest issue.

Enjoy this latest issue of SDSC Innovators!

Michael L. Norman
SDSC Director
SDSC’s ‘Comet’ Supercomputer Enters Early Operations Phase

*Petascale Cluster Open for Researcher Allocations*

Comet, SDSC’s new petascale supercomputer designed to transform advanced scientific computing by expanding access and capacity among traditional as well as non-traditional research domains, has transitioned into an early operations phase.

“Comet is really all about providing high-performance computing to a much larger research community—what we call ‘HPC for the 99 percent’—and serving as a gateway to discovery,” said SDSC Director Michael Norman, the project’s principal investigator. “Comet has been specifically configured to meet the needs of researchers in domains that have not traditionally relied on supercomputers to solve their problems.”

“Based on Comet’s early performance, we are confident in its ability to address a broader set of research topics and communities than past systems,” said Irene Qualters, division director for Advanced Cyberinfrastructure at the NSF. “We congratulate SDSC and UC San Diego on their accomplishment and look forward to Comet’s full deployment.”

*Read more at http://tinyurl.com/pvht2hl*

SDSC’s New Website Is Here!

In addition to a fresh look and audience-friendly layout, the new website highlights the people behind SDSC’s expertise, resources, and services. With revised, intuitive navigation and multiple content-access points, fewer clicks are needed to reach some content that was previously harder to find. Plus, the website’s responsive design provides easier access “on the go” using mobile devices with smaller screens, such as tablets and cell phones. See for yourself at www.sdsc.edu

During the coming weeks and months the website team will add more features, including a dynamic Projects and Experts database, Q&A profiles, and short videos of many of our PIs as well as tighter integration with our social media sites to spread the word about the wealth of research opportunities, services, and events at SDSC.

Kudos to SDSCers Ben Tolo, Cynthia Lee, Csilla Csori, and Mike Dwyer, who very competently brought this phase of the project to the point of launch.
Rearranging the Scientific Web:
SeedMe Resource Aims to Change How Researchers Share Data

Late last year SDSC received a three-year, $1.3 million award from the National Science Foundation to develop a web-based resource that lets scientists seamlessly share and access preliminary results and transient data on a variety of platforms, including mobile devices. The program is being led by SDSC’s Amit Chourasia, a visualization scientist who lends his expertise to numerous projects. One of SeedMe’s recent presentations was held at the SEA Software Engineering Conference 2015 in Boulder, CO, April 13-17th.

This article was written by Olivia Palid and Sally Yen, who participated in last year’s Research Experience for High School (REHS) student program in which students were paired with SDSC researchers and staff to work on a wide range of computational projects. They wrote this article as part of their internship, which focused on how to write engaging and compelling science stories. Read more about SDSC’s REHS program at http://tinyurl.com/krrht9d

With every sunrise, science is rearranged in some way from the day before. It may be imperceptible to all but the most specialized scientists—maybe a new species of harmless amoeba splits in two—or it may affect every human on the earth—the elusive Higgs-Boson particle is finally found. For professional scientists or science aficionados, it seems impossible to be up to date on every discovery.

Because of this, Amit Chourasia, a visualization scientist at SDSC, has embarked on a globally significant task: developing a safe, secure, and easy-to-use way to collect and share data instantaneously within the scientific community. This web-based resource, known as SeedMe, aims to an indelible imprint on the way research is stored and shared.

Chourasia has been interested in computing and graphics since his childhood in India. However, his low college entrance exam score prevented him from initially pursuing this interest. Instead, he chose to major in architecture, but a grueling internship in Delhi proved that a career in this field might not be his best choice. At the same time, Chourasia’s interest in computers grew because of the technology utilized in his courses. This, along

This image and related research data, one of numerous projects being shared and stored using SeedMe, shows a simple model of geodynamo used for benchmark codes. The view is from center toward one of the poles, and the cones show convective flow toward higher temperature (light green to dark green with increasing velocity) in a spiraling form caused by rotation. The shells of various colors depict temperature, increasing from the outer boundary towards the interior. Courtesy of Amit Chourasia1, Ashley Willis1, Maggie Avery1, Chris Davies1, 2, Catherine Constable2, David Gubbins3, 1UC San Diego, 2University of Sheffield, 3University of Leeds, UK.
with his hobby of programming with a friend, stoked his interest in graphics.

Upon graduation, Chourasia decided to follow his passion by seeking a Master of Science degree in computer graphics at Purdue University. There he honed his skills by helping researchers simulate and visualize the terrorist attack on the Pentagon on September 1, 2001. After achieving his degree, he decided to forego a Ph.D. and apply for a position at SDSC.

One key challenge that Chourasia and his colleagues face is the cumbersome process of disseminating data. This is especially true for creators of scientific visualizations, which have become a larger part of the research landscape in recent years due to advances in computation.

“Since there was too much data to share via e-mail, we would always have to create a webpage and send them a link,” he said. “We had to do this over and over again, so we wanted to find a way to speed up that process.”

Inspired to fix this repetitive, time-consuming process, Chourasia developed the concept for SeedMe. In 2012, he received funds from the National Science Foundation to complete the project. He calls SeedMe a “collaboration vehicle”, in which scientists and researchers can easily share their findings with peers. This site provides the ability to upload simulation data directly from a high-performance computer, set privacy settings, and show instantaneous updates and notifications from simulations.

Perhaps most novel about SeedMe is that information uploaded can potentially be accessed by anyone around the world, whether they are a student in San Diego, a congressman in Washington D.C., or a researcher in London. All of SeedMe’s features are designed to streamline the process to help it become the one-stop-shop for scientific data.

A common stereotype of scientists is that they are a secretive bunch who prefer to keep their findings close to their vests until (or even some time after) they have had a significant breakthrough. Chourasia addresses potential privacy concerns by encrypting and transporting all communication on SeedMe through a secure port. In addition, content sharing can be set to various levels: private, within a pre-designated group, or publicly with no limits. Depending on which level users choose, scientists’ experience on SeedMe may vary from personal to global.

Chourasia addresses any fears that outsiders will steal users’ hard work: “SeedMe provides an ability for users to add credits and license information to their content as they see fit.” Additionally, the site will soon have a Wikipedia-esque flagging system whereby users can report inappropriate content.

In our shrinking world, we are all connected by convenient transportation and pervasive social media. Through shared findings and newfound discoveries, Chourasia’s new network may be just what is needed to weave a larger web of knowledge while drawing the realm of science ever closer. In the process of coming together, as scientists or as a society, we are expanding our horizons, exploring the undiscovered mysteries that this world has to offer, and rearranging the world.

Read more about SeedMe at www.seedme.org
Frank Würthwein, a noted expert in high-energy particle physics and advanced computation, has joined SDSC as head of the Center’s Distributed High-Throughput Computing Group to develop and deploy a high-capacity data cyberinfrastructure across all UC campuses.

A UC San Diego physics professor since 2003, Würthwein was recently named executive director of the Open Science Grid (OSG) project, a multi-disciplinary research partnership funded by the U.S. Department of Energy and National Science Foundation. He was OSG’s founding executive during 2005.

Würthwein is no stranger to processing extremely large data sets. In 2013 he and his team used SDSC’s Gordon supercomputer to provide auxiliary computing capacity to the OSG by processing massive data sets generated by the Compact Muon Solenoid (CMS), one of two particle detectors at the Large Hadron Collider (LHC) near CERN, Switzerland. SDSC’s Gordon is part of the NSF’s XSEDE (eXtreme Science and Engineering Discovery Environment) program.

“Frank’s appointment will help create a seamless interface between the nation’s two leading open scientific computing infrastructures—OSG and XSEDE—which will directly benefit a broad spectrum of researchers,” said SDSC Director Michael Norman. “His expertise paves the way for him to pioneer a shared data and compute platform across the entire UC system anchored at SDSC. Frank’s appointment is just one of many ways we are committed to strengthening our UC engagement efforts.”

One of the key benefits of such a network is that individual PIs across all UC campuses will have direct access to SDSC’s expertise and resources from their home institutions. “We view this network as a key solution and enabler for data-enabled research, and we can see the day when other university systems and research enterprises follow suit with similar systems,” added Norman.

Read more at http://tinyurl.com/p5lfrrm

UC San Diego Physicist Frank Würthwein Joins SDSC

SDSC has announced its inaugural UC Graduate Student Summer Fellowship program, providing opportunities for graduate students throughout the UC system to learn about SDSC’s expertise and utilize the Center’s wide range of resources to advance their own research.

The eight-week residential program, to run from June 22 until August 14, 2015, is funded by SDSC to specifically foster stronger ties with other UC campuses. The program is focused on attracting a small number of graduate students from other UC campuses.

“This program is designed to increase awareness of the value of computational science among the other UC campuses,” said Diane Baxter, SDSC’s associate director for education. “These graduate student fellows will also gain exposure to a more diverse range of career options, gain hands-on computational experience, and add computational research scientists as essential mentors who will help them succeed in their careers.”

Read more at http://tinyurl.com/pceu2qy

“Ada’s Lunch Club” Debuts at SDSC

SDSC has announced a new outreach program for UC San Diego undergraduate and graduate students majoring in computer science, electrical engineering, and biotechnology. The program, named for Ada Lovelace, widely recognized as the first computer programmer for authoring the first algorithm to be performed by a machine, will host a small group of students for an informal monthly lunch with one of SDSC’s female scientists. As each one of our scientists have varied backgrounds in fields such as data mining, IT Systems, or Internet research, the goal is to bring together students with similar interests and career goals to encourage, guide, and mentor. For more information please email amason@ucsd.edu or contact her to receive the new Ada’s Lunch Club e-newsletter.

Read more at http://tinyurl.com/p5lfrrm
What parts of your career have been the most fun?
Pfeiffer: In my 30s, I did nuclear fusion research. The physics problems were really challenging, in fact so challenging that eventually I decided to do something more practical. From my experience in fusion, I had remotely used supercomputers at Lawrence Livermore National Laboratory, and helped Sid Karin write a proposal to found the Supercomputer Center here in San Diego. I stepped down from line management over 10 years ago and in recent years have been doing bioinformatics research.

How did you get into bioinformatics?
Pfeiffer: Six years ago I learned that SDSC Biologist Mark Miller developed software that let other biologists do phylogenetic analyses of DNA sequence data on an SDSC computer cluster via a browser interface rather than logging onto the cluster. The CIPRES portal, or gateway looked promising but each analysis was running on only a single core. To broaden the gateway’s appeal, I helped implement the analyses done by the three most popular phylogenetic codes on more powerful clusters at our sister supercomputer centers in Illinois and Texas, where each analysis could run in parallel on several cores simultaneously. This allowed more computationally demanding analyses and significantly increased usage of the gateway.

Next I developed a hybrid parallel version of one of the codes, called RAxML, which allowed it to run on tens of cores at a time and further improved its performance. This was done via an e-mail collaboration with Alexis Stamatakis, lead developer of the code in Germany. Soon afterward, SDSC launched Trestles and subsequently Gordon, allowing us to move the CIPRES gateway analyses back to SDSC from Illinois and Texas. I have since helped make several more phylogenetic analysis codes available on our supercomputers. CIPRES has been much more successful than Mark and I ever imagined.

Could you describe some of your recent bioinformatics projects?
Pfeiffer: More than three years ago I joined a collaboration led by Henne Holstege of VU Amsterdam in which we used whole-genome sequencing to determine how many somatic mutations had accumulated in the DNA of a Dutch woman over her 115-year life. We looked for mutations in her white blood cells that were not present in her brain cells immediately after she died and reasoned that those mutations had built up from the much larger number of cell divisions in her blood as compared to her brain.

Based upon our analysis we estimated that our subject, whom we called W115, had about 450 somatic mutations in the non-repetitive genome of the white blood cells studied, corresponding to an average of four mutations per year. However, we found only four mutations that mapped to regions in genes that code for proteins, whereas most were in genomic regions predicted to have neither adverse nor favorable impact on genetic fitness. The message here is that one can have lots of somatic mutations and a long life, provided the mutations do not affect genetic fitness. Our results were published last year in Genome Research.

Currently I am collaborating with Josephine Braun and Carmel Witte of the San Diego Zoo on a project to determine whether birds in the Zoo’s collection obtain Mycobacteria infections from each other or from the environment. This involves looking for mutations in each bacterial sample that are not present in the others using a software pipeline that I developed and run on Gordon.

Any thoughts about your future at SDSC?
Pfeiffer: SDSC is as exciting a place to work as it ever has been, which is why I am still here. The Center remains a magnet for researchers needing access to powerful computational resources, and there are many good years still ahead!
At its core, the Big Data phenomenon is about the ability to collect, store, and analyze massive amounts of data—to facilitate new discoveries in the case of science, improve policymaking in the case of government, and achieve a competitive edge in the case of industry. Increasingly, the process of analyzing extreme data sets is accomplished through "pipelines"—passing data through a series of processing steps often involving different software applications and computing resources at each step. As data pipelines become more complex, involving multiple software applications and heterogeneous computing platforms, there is a need for tools to orchestrate the resources, marshal data through the pipeline, and allow for automation, repeatability, and reproducibility of results. This is where the emerging class of tools called Workflow Automation Software (WAS) comes in.

SDSC’s Workflows for Data Science Center of Excellence (WorDS), under the leadership of Ilkay Altintas, is pioneering this field of WAS. If your company is in the process of implementing big data analytics pipelines, whether in the field of bioinformatics, sensor networks, smart manufacturing, or elsewhere, we encourage you to contact the WorDS team and consider a collaboration.

Ron Hawkins
SDSC Director of Industry Relations

WorDS
Housed at SDSC, the WorDS Center of Excellence is a hub for the development, promotion, and delivery of workflow services for a wide range of applications. Its mission is to support data analysis projects, data scientists, and software engineers in their computational practices involving process management.

Expertise and services:
- Consulting with world-class researchers and an A-Team of developers well-versed in data science and scientific computing technologies
- Collaborative development of the popular Kepler Scientific Workflow System
- Development of data science workflow applications through a combination of tools, technologies and best practices
- Hands-on consulting on workflow technologies for big data and cloud systems, i.e., MapReduce, Hadoop, Yarn, Cascading
- Technology briefings/classes on end-to-end support for data science workflows

Areas in which WorDS researchers conduct industrial collaborations include:
- Bioinformatics and biomedical big data
- Cloud systems for smart manufacturing
- Educational big data analytics
- Computational chemistry
- Computer-aided drug design