DIRECTOR'S LETTER

SDSC's Response to the "BIG DATA" CHALLENGE

During the past couple of years, our planet began confronting what many view as a tidal wave of information stemming from our academic centers, commercial laboratories, government scientists, and observational tools such as satellites, oceanographic sensors, astronomical observatories, and personal websites. The term "Big Data", used widely to describe this phenomenon, suddenly became a major topic across a wide spectrum of interests and disciplines which began seeking ways to tame, harness, and otherwise glean knowledge from this ever-widening deluge of data.

Thought leaders from the highest levels of the federal government to board chairmen at the world's leading private and public enterprises have met in workshops, seminars, and other local and national gatherings to better understand the concept of Big Data, initially, at least, to separate hype from reality. These leaders hope the ongoing Big Data conversation yields better narratives and/or visualizations needed 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.

Our nation's research universities, a traditional source of innovation and discovery, also have been brought into the discussion by the business community and government leaders. Here at UC San Diego we're being challenged to develop new tools to better mine, model, manage, and otherwise analyze all this data; to build effective methods to store and curate huge volumes of information over long periods of time; and to develop curricula to help educate and train experts to fill a rapidly growing need for experts in all things relevant to Big Data. The term "data science and engineering" has emerged as academia's answer to the Big Data need, much the way "computational science and engineering" programs sprung up a couple of decades ago in response to an advancement known as supercomputing.

In this year's annual report, we describe an innovative program at SDSC called the Institute for Data Science and Engineering (IDSE), which provides a roadmap for UC San Diego's response to the Big Data challenge. As outlined, IDSE will host and coordinate undergraduate and graduate education and training of Big Data experts at UC San Diego, while also providing a focal point for research collaborations across campus, particularly for its recently unveiled strategic themes that include: understanding and protecting the planet, en-

riching human life and society, exploring the basis for human knowledge and creativity, and understanding cultures and improving societies.

IDSE also will offer an infrastructure to support Big Data such as: scientific simulation and visualization, data modeling and integration, database design and implementation, scientific workflow automation, data mining and predictive analytics, and management of protected data. Over the course of its nearly three-decade history, SDSC has fostered research collaborations and partnerships across a variety of disciplines and departments at UC San Diego and beyond, and we envision IDSE as a lightning rod for new efforts to advance this campus' national influence in data science and engineering.

The evolution of data science and engineering complements SDSC's more traditional focus on advancing computational 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. For example, next-generation sequencing of DNA and RNA—the foundation for the advancing era of "personalized medicine"—is creating a deluge of data that requires increasingly powerful supercomputers to process, and the need to develop complex software to more rapidly and efficiently analyze. SDSC is leveraging its Big Data resources including the *Gordon* data-intensive supercomputer to support efforts in genomics and bioinformatics, working with researchers at UC San Diego and UC, neighboring non-profit research centers on the Torrey Pines mesa and beyond, and California biotech companies.

This year's annual report shines a spotlight on SDSC's vast array of Big Data and computational science capabilities that includes its supercomputing resources, advanced networking and data storage infrastructure, along with brief descriptions of this Center's local impact, statewide influence, and national reach in education, training, and research. The report also excerpts significant research advances made by, or with the help of, SDSC staff and resources during the past year that include new academic collaborations and industry partnerships.

Over its nearly three-decade history, SDSC has garnered a national reputation for excellence in data and supercomputing and, partly as testimonial to this reputation, was awarded a $12 million grant during 2013 from the National Science Foundation (NSF) to build a next-generation supercomputer capable of one quadrillion arithmetic calculations per second. This petascale data-intensive supercomputer is fittingly called *Comet*, connoting its mission of catering to what's being called the "long tail of science," whose goal is to address the needs of a large number of small- to medium-sized computationally based research projects nationwide. In essence, *Comet* is all about high-performance computing for the 99 percent, and is designed to deliver a significantly increased level of computing capacity and customizability to support data-enabled science and engineering for the nation. *Comet* will have all the features that made SDSC's *Trestles* computer cluster so popular with users, and will be particularly well-suited to science gateways that serve large communities of users, such as those new to XSEDE (NSF's eXtreme Science and Engineering Discovery Environment) which comprises the most advanced collection of integrated digital resources and services in the world.

We realize that the computational and data science resources at SDSC are only as good as the people who use them. A critical mass of expertise across numerous specialties support and integrate these various components into a holistic working environment. These individuals include scientists with Ph.D degrees in computer or domain sciences who, through their collaborations, have become cross-trained in a range of computational/domain sciences to help solve problems in the latter. An emerging description of such people is "pi-shaped," with one leg in information technology, one leg in a scientific domain, and a bar across the top indicating the know-how to bridge the two culturally distinct communities.

SDSC is proud of its cadre of "pi-shaped" researchers and, for the first time, we have chosen SDSC's Ilkay Altintas as SDSC's "Pi Person of the Year." Please read about Dr. Altintas' accomplishments later in this annual report, and learn more about how SDSC is shaping its future, and those of its partners and collaborators, in this exciting new era in science and technology.

Michael L. Norman
SDSC Director