UC San Diego Data Analytics Initiative Working Group: Statement of Principles

Terrence August, Ginny De Sa, Erin Glass, Gordon Hanson, Michael Holst, Rob Knight, Mike Norman, Mohan Paturi, Molly Roberts, Larry Smarr, Frank Vernon

November 28, 2016

Advances in sensing and computing technologies are enabling the capture of data at unprecedented levels of detail, variety, spatial and temporal scale, and cost effectiveness. Data-driven modelling offers tremendous opportunities for advancing our understanding of real-world phenomena and improving our capacity to address global challenges, ranging from mapping the brain to managing climate change. Although business, government, and the academy have nominally embraced data science, critical obstacles need to be overcome before we realize the full potential of big data. Having met on three occasions, the data analytics working group has agreed on a broad approach for how to advance data science at UC San Diego. Our objective is to help campus establish itself at the frontier of research both in basic data science and its applications in domain sciences. In this document, we lay out a set of founding principles for the creation of a data science initiative on campus.

Our group did not reach consensus on the organizational form that the data science initiative should take (other than that it not be an ORU). Some thought we should move quickly to establish a Data Science Institute, that would stand free of any individual department and would include a small nucleus of faculty (e.g., 3-4 FTEs) whose primary appointment resides in the entity and a larger group of faculty (e.g., 12-15 FTEs) with a secondary appointment in the institute and a primary appointment elsewhere on campus. Others thought that we should instead follow the model of Integrated Digital Infrastructure, which would entail creating a virtual entity to manage data science on campus. To create this entity, the Chancellor would convene a panel with a short lifespan to bring key constituencies together, to identify gaps in our capabilities, and to design universal approaches to fill these gaps. We mention these alternatives at the outset, without declaring which we collectively agree to be the most promising.

1. *Organize the data science initiative around specific problems and applications*

UC San Diego has the potential to be recognized as a global leader in data science. Our campus has an unusual combination of strengths in data analytics, which derive from our computing infrastructure (SDSC, Catlit2), capabilities in machine learning and applied statistics (APM, CSE, Cog Sci, ECE), and ease of forming multidisciplinary ventures (e.g., the Microbiome and Microbial Sciences Initiative, the Center for Brain Activity Mapping, Institute for Neural Computation, Temporal Dynamics of Learning Center, Institute of Engineering in Medicine, Contextual Robotics Institute, Center for Wearable Sensors). Yet, we have not achieved broader recognition for these strengths. To merit such recognition, we need to present ourselves as solving critical problems in the field of data science while at the same using data analytics to address pressing global needs. Our first step should thus be to identify the specific sets of problems we wish to tackle and the signature applications we see as best communicating our

potential. These problems and applications would help UC San Diego create a brand in data science. We list possible problems and applications in an appendix to this document.

Identifying a compelling set of problems and applications would also help with attracting extramural support. The donor community is likely to embody diverse interests regarding data analytics. Some donors will care not about the methods but about the questions we are trying to answer (understanding the brain, mapping the human biome, protecting the environment, alleviating poverty). Other donors may be attracted by our potential to expand the frontier of data science. Yet others may be concerned about how we train our students and improve their career options. It is important that we present our efforts in data science in an ecumenical manner that simultaneously attracts those with issue-specific concerns, those entranced by the methods, and those whose main concern is student training. By jointly embracing data-science problems and applications, we will create multiple portals through which potential donors could engage with us. Attracting funding may also require that we forge partnerships with private companies that control or manage the raw data that we use in our research, which will require that we think carefully about privacy concerns and how to identify campus contributions to the creation intellectual property that these partnerships would produce.

2. *Incubate new ideas for data science*

Our true strength in data science lies not in the projects that we have already launched, but in our potential for mining our intellectual diversity to achieve major new breakthroughs. These breakthroughs require that the data science initiative develop strategies for connecting researchers who have strong intellectual affinities but may not know it yet. A major function of the initiative should thus be to create the connective tissue that will bring researchers from disparate parts of campus together to form collaborations. Geography is an inhibitor in getting scholars together. If we are to incubate new ideas effectively, we must find mechanisms that attract fertile minds to gather on a regular basis and in uncommon groupings. Engagement by UC San Diego faculty in interdisciplinary research is less common than desired. Campus leadership should identify and scrutinize the barriers that are currently throttling throughput of interdisciplinary scientific discoveries. Heterogeneity in departmental governance and evaluation of research output is another inhibitor. It is critical to determine appropriate incentive structures that empowers the typical faculty member to engage in these activities.

3. *Create viable career paths in data science*

There is enormous heterogeneity on campus in how data science is developed and applied. Despite this heterogeneity, our researchers face a common set of personnel constraints, which are preventing UC San Diego from realizing its full data-science potential. Front-end research requires substantial investments in assembling, managing, and preparing data. Data analysis requires skills in machine learning, remote sensing, GIS, crowd sourcing, and other tools. The scarcity of research staff with data management and data analytic skills is holding back our collective efforts. Given the commonality of needs in data analytics, there is substantial scope for

sharing research staff. Campus should work to relax personnel constraints by building a multi-disciplinary team with expertise in data analytics that would include project and research scientists, as well as post-doctoral fellows and new tenure-track faculty. It is highly likely that many of our hires in data science will be researchers with essential skills who nonetheless find themselves as "middle authors" on resulting publications. Campus has had difficulty in rewarding researchers who fill these technical support roles. If we are to expand research personnel in data science, we need to ensure that we have effective means of rewarding the contributions of these individuals. Existing incentive structures on campus are deficient.

Appendix: Problems and Applications in Data Science

(A) Examples of Problems in Data Science

- *Data Life Cycle Management*

Modeling and analysis typically receive the most attention in big data projects. However, all stages of the data life cycle are critical for success, which makes improving productivity at each stage a primary research challenge. The data life cycle includes:

a. Data acquisition,
b. Information extraction and cleaning,
c. Data integration, aggregation and representation,
d. Modeling and analysis,
e. Interpretation, and
f. Archiving, and reuse.

Effective management of the data life cycle would help achieve the additional goals of establishing a basis for scientific reproducibility while maintaining data privacy. As the value of data is increasingly recognized, it becomes part of the strategic calculation of an organization. Questions regarding the sharing of data without renouncing the unique advantages of data ownership will come into sharper focus.

- *Theory, Algorithms and Computation*

At the core of deriving value from data is a theoretical framework for learning from data, modeling data, creating algorithms for effective inference and prediction, and computational efficiency for dealing with large data sets. Advances in these areas are critical for dealing with the inconsistency, incompleteness, and heterogeneity of data.

- *Organizational Challenges*

Big data methods and techniques are deployed by scientists across a wide variety of academic disciplines to deal with consequential scientific, social and intellectual questions. There is a great

deal of commonality in scientists' needs in terms of data management, modeling, and analysis. However, we face the need to lower the intellectual, cultural and organizational barriers so scientists can effectively apply data science methods and techniques for their projects.

(B) Examples of Data Science Applications

- *Mapping the Human Biome*

Each of us has more microbial cells than human cells in our bodies. The million-fold decline in the cost of DNA sequencing over the past decade has accelerated our ability to collect the sequences that characterize these microbes, but at the expense of our ability to understand them. We need new techniques in data science to understand the high level of variation across individuals in their microbiomes and how this variability links to human health. Spatial mapping of the human lung in cystic fibrosis is an example of multidisciplinary research on the microbiome where better data science is needed. Although cystic fibrosis is thought of as a human genetic disease, the difference between living to age 2 (the average 50 years ago) and age 52 (the average today) depends on controlling complex microbial biofilms that form in the lung because of genetic defects. In a multidisciplinary collaboration at UC San Diego (spanning surgery, pulmonology, microbiome, metabolomics, viromics, medicine, ecology, and computer science), explant lungs from cystic-fibrosis patients are diced into small cubes. Each cube is then processed for DNA sequencing and untargeted mass spectrometry to understand the microbiome and metabolome. These highly multivariate estimates covering thousands to tens of thousands of features with intensities in each sample are then related back to the CT or MRI imaging of the same patient's lung before surgery. The spatial patterns uncovered allow us to identify interactions among different microbes, and breakdown products of different drugs, where the specific microbes in each patient's lungs break down antibiotics in different ways, altering which other microbes they can target. We hope to apply the results to modeling sputum so that we can understand these relationships, and treat patients, in cases where the lung has not yet been removed. This requires processing of terabytes of image, DNA sequence and mass spectrum data, fluid dynamic and genome-scale modeling of the lung, and uncovering novel pathways of xenobiotic metabolism, as well as better understanding viral-bacterial ecological interactions.

- *Modelling Wildfire*

Over the past 15 years there has been an explosion of sensor network data from many scientific disciplines (including satellites in space observing magnetic fields and solar wind; meteorological networks for real-time forecasting and climate research; geophysical observations of earthquakes and tectonic motion; physical oceanographic measurements of currents, temperature/salinity, waves and acoustic tomography/thermometry; and cameras for detecting and observing wildfires). The number sensors where data can be acquired in scalar time series, the precision of the data, and the dimensionality of the data have all increased. In the early 1970s, a continuous series of a million samples was rare. Such series are now on the small side. With the advent of virtually ubiquitous networking, environmental sensor data are being

continuously streamed to a variety of locations for immediate application (e.g., tsunami detection, earthquake early warning). One example is the High Performance Wireless Research and Education Network (HPWREN), which provides research, education and public safety support over its San Diego County wireless network. HPWREN provides access to hundreds of sensors and instruments to the scientific community, network access to regional back county fire stations, and real-time public access to a multitude of cameras across the county via the internet. HPWREN cameras and meteorological sensors are a prime example of data science applications in environmental studies. HPWREN provides streaming data used for detection and observation of wild fires, as well as meteorological data that inform near real time fire modeling, visualization, and data assimilation from the UC San Diego Wildfire Project. Public real time access to these data are a critical aspect of these systems to provide decision makers and the general public the ability to make informed decisions in a timely manner.

- *Measuring the Political Climate in China*

The opaque nature of China's political system prevents outsiders from witnessing domestic political dynamics and creates uncertainty over China's willingness to exercise its power globally. To better measure Chinese leaders' intentions and directions, the Center for Applied Internet Data Analysis (an independent analysis and research group based at SDSC) and the 21st Century China Center (at GPS, which also involves faculty from DSS) are integrating data from social media, newspapers, biographies of Chinese officials, government documents, and social surveys to provide a new window into China's political climate, society, and economy. The newly created China Data Lab hosts an internal data platform that leverages new data analytic techniques to integrate real-time releases of unstructured data by time, text and geography, which enables efficient cross-referencing between and summaries of political indicators across datasets. The China Data Lab is also developing a public-facing platform to make data, analyses, and workshops on data analysis in China broadly available.

- *Mapping Urban Areas at Global Scale*

Urbanization in China, India, and other rapidly growing nations is helping lift hundreds of millions of people out of poverty. However, it is also creating immense societal challenges by expanding greenhouse gas emissions, destabilizing fragile ecosystems, and upending traditional social networks. Existing approaches to measuring urbanization use infrequent household surveys, which are slow and costly to produce and which governments can easily manipulate. To create sound environmental, health, industrial, and transportation policies for global cities, we need the ability to track urban growth in micro detail and in real time. Through the Big Pixel Initiative, researchers at GPS, Calit2, and SIO are working with Google Earth Engine to apply the power of machine learning to vast new troves of satellite imagery. We aim to understand the causes and consequences of urbanization in order to design public policies that create economic opportunity while promoting environmental stewardship. Our approach involves applying machine-learning algorithms to satellite imagery to describe land use at high spatial and temporal resolution, validating the accuracy of the resulting urbanization measures via crowd-sourcing, and using the data to analyze how industrialization (e.g., the arrival of new large manufacturing

facilities) and infrastructure investment (e.g., the building of new highways high-speed rail systems) affect how cities grow and develop in micro- and macro-geographic detail.

- *Understanding Brain Imaging Data*

Understanding the human brain is one of the great challenges recognized by President Obama in forming the B.R.A.I.N. initiative. The director of the Data Science Institute of Imperial College estimates that "mapping brain activity will produce nearly as much data as the Large Hadron Collider." Scientists in the Swartz Center for Computational Neuroscience (in the Institute for Neural Computation), the Multimodal Imaging Laboratory (MMIL in the Departments of Neuroscience and Radiology) and several labs in the Departments of Biomedical Engineering, Cognitive Science, Computer Science and Engineering, and Electrical and Computer Engineering are developing and applying new methods in signal processing and machine learning to analyze human ECoG, EEG, fMRI, MEG, and MRI data and, in the case of the MMIL, concurrent genomic data. The goal is to better comprehend the computational processing of the human brain for the purposes of understanding the development, structure, and function of the human brain, and their genetic and environmental influences, as well as for creating brain-computer interfaces (directly reading human intention from brain activity). The Center for Brain Activity Mapping has recently funded several projects that increase the capabilities for real-time sensing and recording of neural activity, genetic activity, and neurotransmitter activity in animal and human brains. This will create new forms of multimodal imaging data that will require new advances in data science methods.