UCDAVIS
UNIVERSITY OF CALIFORNIA

BIG DATA IMPLEMENTATION COMMITTEE REPORT

Submitted to Provost Hexter

February 22, 2013

TABLE OF CONTENTS

I. EXECUTIVE SUMMARY ... 3

II. A VISION FOR UC DAVIS ... 5

1. Big Data, Big Challenges, Big Opportunities ... 5
2. UC Davis and Big Data: State of the Art and Current Needs ... 6
3. The UC Davis Data Science Institute ... 7

III. RESEARCH ... 9

1. Big Data Academic Leadership, Research Talent and Support Requirements ... 9
2. Big Data Exploration: A New Paradigm for Discovery ... 9
3. Big Data: Enabling Large-scale Simulations ... 10
4. Big Data Exploration And Knowledge Discovery ... 11
5. Data Governance and the Social Aspects of Big Data ... 13
6. Big Data: Accelerating Innovation at the Interface between Disciplines ... 14

IV. TEACHING ... 16

1. Big Data Analytics Training for Postdoctoral Fellows, Staff, and Faculty ... 16
2. Data Science at the Ph.D. Level ... 17
3. Data Science at the Master’s Level ... 17
4. Data Science at the Undergraduate Level ... 18
5. Outreach ... 18
6. Costs to Implement and Run the DSI’s Education Program ... 18

V. INFRASTRUCTURE ... 19

1. What Our Model Must Encompass ... 20
2. Current State of the UC Davis IT infrastructure ... 20
3. Critical Infrastructure Components to Support Big Data Research ... 21
4. Implementation ... 26

VI. DATA SCIENCE INSTITUTE: PROPOSED IMPLEMENTATION TIMELINE ... 27

VII. APPENDIX A -- VIGNETTES: BIG DATA @ UC DAVIS ... 28

1. Big Data and the Humanities ... 28
2. Big Data in the Social Sciences ... 29
3. Big Data in Cosmology ... 30
4. Big Data and Environmental Sciences: Opportunities ... 30
5. Big Data in Computer Science ... 32
6. Big Data in the Life Sciences: Opportunities and Challenges ... 33
7. Big Data and Precision Medicine ... 34

APPENDIX B: INFRASTRUCTURE AND SERVICES SUPPORTING RESEARCH AT UC DAVIS ... 36

APPENDIX C: BIG DATA INFRASTRUCTURE EFFORTS ELSEWHERE ... 39

APPENDIX D: CHARGE LETTER AND MEMBERSHIP ... 41

APPENDIX E: ACKNOWLEDGMENTS ... 43

I. EXECUTIVE SUMMARY

"Where is the wisdom we have lost in knowledge?
Where is the knowledge we have lost in information?"
Where is the information we have lost in data?
With apologies to T.S. Eliot

The Future Has Arrived

Advances in technology and the ever-growing role of digital sensors and computers in science have led to an exponential growth in the amount and complexity of data that scientists collect. We are at the threshold of an era in which hypothesis-driven science is being complemented with data-driven discovery. This alternative way to pursue research affects all fields, from genomics in biology, to astrophysics and many domains in social sciences. The data collected are complex in size, dimension, and heterogeneity — all three generating what is generically referred to as "Big Data." These data provide unprecedented opportunities for new discoveries; they also come with challenges that must be addressed.

UC Davis Is at the Forefront of the Big Data Wave

The quality, diversity, and successes of research at UC Davis are recognized worldwide. Many of our research programs are engaging the Big Data wave by necessity. The Genome Center is enabling comparative functional genomics and generates terabytes of data every week. The impending application of whole genome sequencing to clinical care will require secure storage of terabytes of data linked to medical records. The observational cosmology group is a leader on the Large Synoptic Survey Telescope project, recently ranked top priority for the next ground-based facility by the National Academy of Sciences; it will soon have to deal with hundreds of petabytes. UC Davis has already developed expertise to support these areas. Also, there are researchers in physics, computer sciences, mathematics, and statistics that specialize in data sciences. IDAV, VIDI, and KeckCAVES are world-leading efforts in scientific data analysis and visualization, and the Universe@UC Davis Initiative is an early example of a successful multi-disciplinary collaboration on the frontier of physics and information involving faculty from several departments as well as collaborators at Google, and scientists at LANL and LLNL.

While these initiatives are of top quality, they are fragmented, with too few resources to cover the needs of the campus. We need to foster synergetic connections between these groups to increase cross-fertilizations as well as increase their sizes such that they can reach the critical mass necessary to connect data with discoveries and understanding.

What Is Needed

We must address the opportunities related to Big Data on three fronts — namely Research, Education, and Infrastructure. To explore scientific frontiers based on Big Data, we must develop novel efficient algorithms for dimensionality/complexity reduction, tools for statistical analysis, and approaches to data exploration and visualization. Just as clever algorithms enabled efficient fast textual search, a new universe of discovery will be opened across our campus if we address these opportunities. We must develop and foster expertise and skills relating to Big Data and Data Sciences in a novel inter-disciplinary teaching environment in which colleagues join together to develop and share solutions. We must develop the campus IT infrastructure to facilitate access to large datasets within a secured environment. Finally, we need to provide a home for these initiatives to develop, one that fosters interactions among disciplines, thereby promoting cutting-edge discoveries and serving as a magnet for researchers on campus and beyond.

If We Do Nothing

It is possible to envisage solutions with limited scopes that partially address the current campus Big Data shortcomings related to Research, IT Infrastructure, and Education. However, any such solutions would seriously constrain UC Davis' ability to support the current and prospective research efforts involving Big Data, enable new programs in areas not yet exposed to Big Data, and train the new generations of students who will have to deal with Big Data. The funding of many research efforts on campus will be contingent on having access to the tools and expertise for analyzing the Big Data they generate. A lack of investment in the domains of Big Data will also be detrimental to the campus' ability to attract excellent new faculty who will expect support to deal with their Big Data. Ultimately, being swamped by the Big Data wave will jeopardize the ability of the campus to remain a leading research and teaching institution.

What We Propose

We propose the creation of a **Data Science Institute** to bring together mathematicians, computer scientists, statisticians and domain scientists to work on frontier research and Big Data challenges in a collaborative setting. The goal is to develop an intellectually stimulating physical environment that will nurture collaboration among scientists who tackle the challenges of Big Data, and develop a space for domain scientists to interact with these experts. This Institute will serve as the nexus for data science technologies for the entire UC Davis campus. Permanent members of the Institute will have duties in Research, Education, and Service related to promoting Data Science to diverse communities on campus.

Implementation

While ultimately the Data Science Institute will need its own physical space, its virtual existence should be an immediate priority. In the next few years, we must:

- Invest in **infrastructure** (bandwidth, accessibility) to facilitate access to data on campus. This will be best implemented by defining pilot programs supported by the campus and implemented within a short time frame;
- Hire several **key senior faculty** now to establish a core membership for the Institute. In particular, we need to identify and hire an inter-disciplinary leader in the field of Data Science who will bring scientific excellence as well as leadership in making the Institute successful;
- Hire several **key young faculty, postdoctoral fellows, and staff scientists** to create a critical mass for the Institute to thrive. The task of the staff scientists will be to develop a core service on campus for Data Analytics. This core service will complement and support the domain capabilities for data analysis already active on campus; and
- Support initiatives for developing **new courses in Data Analytics** for undergraduate and graduate students as well as for off-campus professionals, in the form of certificates.

II. A VISION FOR UC DAVIS

1. Big Data, Big Challenges, Big Opportunities

"Big Data" refers to data sets whose complexities are such that it becomes difficult or even impossible to capture, manage, and/or process them in a reasonable amount of time with the currently available software tools. In this context, the data can be complex in size, dimension, and/or heterogeneity.

The concept of Big Data is not new: physicists for example have been struggling with limitations in memory and processing power from the time they started using computers, and industry has always had to deal with large data warehouses. It is, however, the exponential growth in data complexity and their recent full-fledged democratization that is overwhelming. In 2010, Eric Schmidt, Google CEO, noted that "every two days, we create as much information as we did from the dawn of civilization up until 2003." This explosion in the amount of information comes mostly from the widespread use of mobile devices and digital sensors, the boom of social networks, and the development of international large-scale experimental programs. Whereas scientists had to struggle with Megabytes (10^6 bytes) of data in the 1990s, many datasets in 2013 include many Petabytes (10^{15} bytes) of data, and this number is expected to grow to the Exabyte (10^{18}) and even Yottabyte (10^{24}) scales before 2020 (the National Security Agency is already building a data center that can store data at the yotta scale, that is expected to be operational by the end of 2013). In the rest of the report, we will refer to the level of computing required by Big Data as exascale, alluding to research that requires 10^{18} elements of various types (compute cycles, bytes of storage, bandwidth between computers and for human-computer interaction); exascale computing is the next major milestone for computational performance, but clearly there will be others.

As data are now everywhere, this has given them economic, industrial, and scientific importance. Governments and businesses have recognized that many opportunities stem from these Big Data and have consequently invested heavily in "Big Data Research Initiatives"[1] to meet with the large scientific challenges that need to be addressed before these opportunities become concrete.

The opportunities and consequences of Big Data are far-reaching. In the scientific environment, it is now possible to extract knowledge and understanding at a scale never envisioned before. Jim Gray[2] referred to it as the "Fourth Paradigm" and introduced the concept of data-driven science, expected to complement the more traditional hypothesis-driven science. Big Data opens new frontiers for simulations, making it possible to explore alternative models of nature that can be tested against experimental observations, thereby leading to trustworthy analyses and deeper understanding. Big Data provides the means to move beyond reductionism and develop efforts to study complex systems in a more holistic approach. The Human Brain Project[3], recently funded by the European Union, epitomizes all these opportunities: it aims at simulating the human brain, with the ultimate goal of allowing neuroscientists to connect the dots leading from genes, molecules and cells to human cognition and behavior. Similar opportunities arise in the social sciences from socially generated Big Data, observed as a result of human interactions that are increasingly recorded via web and mobile devices, or revealed through digitization of historical records.

The challenges that Big Data represent are, however, on par with the scale of their opportunities. As a recent (Feb 12, 2013) *MIT Technology Review* reports, the challenge "is not processing or storing this amount of data — Moore's law should take care of all that. Instead, the difficulty is

[1] See for example the "Big Data Initiative" from the Obama administration, announced in 2012: http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf

[2] http://research.microsoft.com/en-us/collaboration/fourthparadigm/

[3] http://www.humanbrainproject.eu

uniquely human. How do humans access and make sense of the exascale data sets?"[4] We must develop new methods for extracting knowledge and understanding from Big Data. Current algorithms for complexity reduction, statistical analysis, and data exploration do not scale beyond the terabyte at best. Data of these magnitudes will also raise severe visualization challenges, as their complexity exceeds the capacity of the brain to extract information from images. We must develop new infrastructures to facilitate access to, and store, these data. Many data centers housing massive storage systems are being developed around the world to solve this problem. There is a parallel need to make the stored data secure, both in terms of reliability of the storage solution and in terms of privacy and confidentiality. This concerns data ownership, usage rights, the right to be forgotten, or ethics. Researchers in the humanities and social sciences must address these issues. Finally, we must address the question of training for Big Data. A recent report by McKinsey[5] estimated that the United States alone face a shortage of 140,000 to 190,000 scientists with data analytics expertise and 1.5 million managers and analysts with the skills to understand and make decisions based on the analysis of Big Data.

2. UC Davis and Big Data: State of the Art and Current Needs

UC Davis has already been impacted by the Big Data wave. Several research efforts on campus are already generating huge amount of data. We describe a few examples of these efforts in vignettes included as Appendix A in this report, in the fields of humanities, social sciences, cosmology, environmental science, genomics, computer science, and precision medicine.

UC Davis already has some expertise to support these research efforts that generate Big Data. Several research groups are world leaders in the fields of simulations and data sciences; they are currently working separately in the departments of Mathematics, Physics, Chemistry, Computer Sciences, Statistics, and various fields related to the Environmental Sciences. There are no faculty members, however, that directly specialize in Big Data. UC Davis has developed world-renowned programs in Data Visualization and Data Analysis, such as IDAV, KeckCAVES, VIDI, and more recently the UC Davis Center for Visualization, sponsored by the UC Davis Research Investments in Science and Engineering (RISE) program. All these efforts, however, are fragmented, with little capacity to cover the needs of the whole campus.

UC Davis needs to rely on a solid IT infrastructure (see Appendix C for an overview of its current status). It includes a high-speed data network that is accessible to all UC Davis faculty, an advanced research network made possible by a recent (summer 2012) NSF grant, a data center, and a set of high performance computing clusters. The joint UC Davis-Health System initiative to replace the old data centers at both campuses with a new 70,000-sq ft facility will equip the University with a much needed environment to develop data storage and meet the Big Data needs of the community in the near future. It is critical, however, for the IT infrastructure to remain well equipped with the most advanced IT solutions to anticipate the needs that are exploding across the disciplines (as clearly called out by all the major funding agencies), but especially over the next several years as UC Davis faculty compete aggressively for funding with peers at other institutions that have already begun to make the needed investments.

UC Davis is already offering courses in data analysis, modeling, and visualization at the undergraduate and graduates levels in the departments of Statistics and Computer Science. Most of these courses, however, are dedicated to students enrolled in these departments and are not accessible to a broader range of students coming from diverse backgrounds. Ad-hoc courses have been developed in the "domain sciences" (Biology, Physics, …); these courses would benefit from being taught by faculty from core Big Data disciplines who are more likely to introduce the most

[4] http://www.technologyreview.com/vi-exascale-computing/

[5] http://www.mckinsey.com/features/big_data

recent techniques relevant to Big Data. Finally, there are few opportunities for staff and faculty on campus to be (re)trained on advanced data analytics.

It is possible to envisage solutions of limited scope that partially address the current shortcomings of the Research, IT Infrastructure, or Education components of Big Data. However, this would seriously constrain UC Davis' ability to not only support current and prospective research efforts that involve Big Data, but also enable new programs in areas not yet exposed to Big Data, and train the new generations of students that will have to deal with Big Data. The funding of many research efforts will be contingent on having access to the tools and expertise for analyzing the Big Data they generate. A lack of investment in the domains of Big Data will also be detrimental to the campus' ability to attract excellent new faculty who will expect support with their Big Data. Ultimately, being swamped by the Big Data wave will jeopardize the ability of the campus to remain a leading research and teaching institution. The campus barely has the capability to deal with its current Big Data needs; it has the ambition to expand in this domain, yet the current research programs and the supporting infrastructure are not as comprehensive as they need to be to remain competitive. UC Davis needs to implement a long-term vision to support, enable, and amplify Big Data research on campus. We therefore propose the creation of a Data Science Institute to bring together researchers from the core disciplines of Big Data and from domain disciplines in a structure that will foster collaboration and cross-fertilization of ideas.

3. The UC Davis Data Science Institute

The mission of the UC Davis Data Science Institute (DSI) is to support and enable Big Data science to accelerate discovery at the frontiers of scientific, engineering and social disciplines.

Ideas and methods from Mathematics, Computer Science, Statistics, and other fields will be brought together to address the problems raised by Big Data and exascale computing in domain sciences. In turn, the domain sciences will create new opportunities in these fields. Students, staff, and faculty from both horizons (Big Data core disciplines and domain sciences) will share a common space that will foster collaboration. They will broaden and enrich their training and be better prepared to approach the Big Data challenges. The Institute will develop and maintain an IT infrastructure for transferring, storing, analyzing, and visualizing Big Data in a secure way that will support not only its research enterprise but also the research of many other UC faculty members.

We envision that the DSI will serve the following purposes:

- Advance the core disciplines of Big Data; enable exploratory data science research (including algorithm innovation) to identify new techniques and technologies that can help advance the field.
- Enable data-intensive domain sciences by making these advances widely available. Significant discoveries will be accelerated through formation of collaborative teams involving data analytics and domain-based efforts.
- Provide outreach to the community; create new opportunities in fields not yet exposed to Big Data.
- Implement education programs and train the next generation of specialists in Data Science, up to Ph.D.-level individuals.
- Deliver education to the broader campus community, developing the next generation of interdisciplinary students and scholars. Educating all students, staff and faculty in the core tools and techniques required to investigate Big Data will result in individuals who are competent in accessing, manipulating and extracting knowledge from massive information sets.
- Generate vocational training via a master's-level degree for information technology professionals and provide them with the know-how to interact with Big Data.

- Serve as the center for information technologies relevant to Big Data for the entire UC Davis campus. The excellence and expertise of the Institute's core faculty will position them well to anticipate technology needs earlier than most individual research groups could.

Key to the Institute's success will be the scientific excellence of its leadership and core members, as well as its visibility on campus and acceptance by the UC Davis community. To maximize its impact in Research, Education, and Infrastructure, we recommend the following program components.

i) Creation and management of hotel space in the DSI

The DSI provides a space where scientists from the core disciplines of Big Data sciences and domain sciences interact and have access to the tools they need to succeed. Access to this environment and to these facilities should not be limited, however, to the core members of the Institute. Instead, reserving space will be possible for affiliates with needs that may not initially justify full DSI membership. This "hotel space" affords opportunities for researchers to collaborate with experts in data science during the early stages of projects, and provides options to "test run" ideas still in their infancy or build new collaborations. Hotel space may be used specifically to get access to a facility (such as visualization platforms). Occupancy will be limited to one year to keep the space dynamic. All UC faculty and researchers will be eligible to apply for hotel space. This space will be allocated by a supervising committee whose members are drawn from the DSI core and affiliated faculty.

ii) Creation of a data science core service

It is critical to the success of the UC Davis Big Data initiative to combine a 'campus service' element with a 'data science research' function. Through such juxtaposition we can ensure that the campus stays competitive by constantly advancing the state of the art in data science and related disciplines, and by rapidly deploying new tools and techniques to a wide variety of data-intensive fields. Professional computer scientists, statisticians, and engineers, from undergraduate to Ph.D. students, will staff this data science core service and provide "routine" support to domain scientists. Due to its generality in data science, this core facility is designed to complement and support the existing domain core services on campus. Finding appropriate job titles for one- or two-year appointments at the Bachelor's level will be important; these jobs could be an excellent step on the career path through graduate school.

iii) The DSI Investment in Interdisciplinary Programs (DIIP)

The DSI aims to enable data science by fostering collaboration between core data science and domain scientists and giving them the tools they need to succeed. We believe that the strongest of these tools will be the DSI Investment in Interdisciplinary Program, or DIIP. DIIP will award seed funding for high-risk, high-reward, collaborative projects across the university. Preference will be given to projects that cannot be funded by other sources, either because they are still in their infancy or because they are too high risk. Preference will also be given to projects that expand beyond the traditional domains of data science, especially social sciences and humanities. It is expected that the DIIP program will increase the nature and diversity of collaboration among faculty across the UC Davis campus.

iv) The DSI Investment in Education

The DSI will play a leadership role in data science education at all levels on campus, providing training in data science to postdoctoral fellows, staff, and faculty in the form of lecture series, workshops, and boot camps; establishing graduate programs in data analytics, both at the Ph.D. level and Master's level; and driving undergraduate education and training. We will seek advice from the academic council and departments on how to implement this component.

The next three sections of this report cover the specifics of the DSI implementation in the areas of Research, Teaching, and Infrastructure. We conclude with a rough timeline that summarizes the steps to be taken for a successful implementation.

III. RESEARCH

The proposed mission of the UC Davis Data Science Institute (DSI) is to develop, support and enable data science across disciplines on campus. It is expected that the Institute will advance scholarship in the core disciplines that underpin Big Data and at the same time provide advanced analytical insight across disciplines. Its implementation, development, and ultimately success will depend critically on its academic leadership; in this section we describe our proposal to develop this leadership. First, we summarize the new paradigms that Big Data is leading to as well as the way they impact experimental, data-rich research as well as computational, simulation-based efforts. We then explore current technologies for exploring data, and how these technologies need to evolve to match the challenges related to Big Data. We describe issues related to data governance, and how they affect social sciences. Finally, we emphasize how Big Data can accelerate innovation at the interface between disciplines.

1. Big Data Academic Leadership, Research Talent and Support Requirements

We propose a 'cluster hire' be pursued to attract a renowned data scientist to act as Director of the proposed Institute, combined with the recruitment of a number of 'translational Big Data experts.' We suggest that the Institute be led by a visionary academic, most likely a recognized leader in a data-focused discipline such as computer science, computer engineering, physics, mathematics or statistics. The lead scientist may work in a research area closely linked to advances in the Big Data field (such as tensor-based computation, massively parallel-processing databases, or general purpose graphical processor units computing), or from a background of scholarship that deeply underpins advances in data science (such as Bayesian statistics), or perhaps new experimental computing paradigms. The data scientist will provide leadership in advancing state of the art in the Big Data field, and concurrently help support and coordinate a network of 'Big Data domain experts' – faculty who specialize in Big Data challenges in specific domains reflecting campus interest and investments. Ideally, the expertise hired across perhaps five or six areas will span the intellectual breadth of the campus. The DSI Director will also closely interact with the campus CIO to ensure the very latest tools and technologies are deployed to the campus community in a timely fashion.

It is critical to the success of the UC Davis Big Data initiative to combine a 'campus service' element with a 'data science research' function. Through such juxtaposition, we can ensure the campus stays competitive over many years by constantly advancing the state of the art in data sciences and related disciplines, and rapidly deploying new tools and techniques in a timely manner to a wide variety of data-intensive fields. We therefore propose that the DSI hosts a Big Data 'analytical techniques' core – a facility staffed by professional computer science/engineering professionals who can provide routine 'Big Data processing' services that do not require 'domain expert' faculty involvement or intervention.

2. Big Data Exploration: A New Paradigm for Discovery

The emergence of Big Data is adding to the repertoire of ways in which research is conducted and discoveries are made. A new approach is emerging in which data – its collection, analysis, and exploration – is driving research. It is critical to note that one should not view data-driven approaches as replacing other forms of research, but rather that they are yet another general approach to understanding the world around us. An excellent summary of this concept is presented in the book "The Fourth Paradigm" based on the ideas of Jim Gray[6]. Gray outlined four paradigms

[6] Hey et al. *The Fourth Paradigm: Data-Intensive Scientific Discovery*, http://research.microsoft.com/en-us/collaboration/fourthparadigm/

for research: empirical (describing natural phenomena), theoretical (using models, generalizations), computational (simulating natural phenomena), and data exploration (which he referred to as the Fourth Paradigm). In essence, one can view the "data exploration" approach as "Let the data speak for themselves."

Perhaps the most critical aspect of "data exploration" relates to large/massive data sets; individuals cannot simply browse through the data and make direct discoveries. Instead, one needs to look for large-scale patterns, correlations, and trends and use statistical measures and other analytical techniques to screen for informative patterns. To be effective in Big Data exploration, we need advanced data management and storage systems (e.g., SciDB), novel data analysis and visualization techniques (e.g., tools to make automated discoveries, methods for dimension reduction and presenting useful visualizations of patterns), and scientific collaboration tools and environments (e.g., to improve interdisciplinary research). Such a data-driven approach complements (and does not by any means replace) "hypothesis testing."

Data driven approaches have begun to take form in diverse areas like genomics, astronomy, business, and history. Importantly, such approaches can be applied to data from experiments, direct natural observations, and massive simulations. And perhaps most importantly, if we truly let the data speak for themselves, we have the potential to make discoveries of the unexpected. In many fields, research advances both linearly and discontinuously; researchers sometimes stumble across an unexpected signal. Traditionally, such disruptive advances have been limited by the number of human minds searching the data. Sometimes curiosity over an unexpected correlation (if not ignored) leads to discovery. This inefficient labor-intensive approach does not scale to the exascale. With advanced database structures coupled with sufficient compute capability, automated discovery of the unexpected is becoming a reality. In addition, improved use of data-driven research can complement other approaches (e.g., it can allow broad testing of hypotheses on a scale unimaginable only a few years ago).

3. Big Data: Enabling Large-scale Simulations

While it is conventional to associate the so-called "data deluge" with the exponential increase in data acquisition observed in experimental sciences, there is a parallel process in science which is generating even more data, namely simulations. With the advent of computers, simulations have become a formidable tool for scientific investigations that tackle nature at its multiple scales. Simulations are designed to remove bias in the interpretation of experimental observations, thus hopefully leading to trustworthy analysis and deeper understanding. More importantly, they provide a framework for connecting the dots between fragmented observations at multiple scales. Simulations can be predictive (such as those used for analyzing the effects of climate changes), they are essential for "what if" studies (for example, to investigate the effects of using pathogens to control insect populations), and they are often used in place of experiments when the latter are too expensive, too dangerous, too time consuming, or even not feasible (when investigating, for example, chemistry at the femto-second to nanosecond time scales). Ironically, simulations can create more of a Big Data challenge than the experimental sciences they complement or replace. A typical simulation in climate sciences, for example, generates hundreds of terabytes of data, and this number will grow to exabytes in the coming years. Simulations of cosmology are producing petabytes.

The developments and successes of simulations have paralleled the improvements in computing technologies over the last fifty years. Scientists in many fields have converted the exponential growth in the performance of computer hardware into the ability to simulate larger, more complex systems over larger time and/or spatial scales. The fastest supercomputers today include millions of processors and perform in the petaflop range (17.6 for Titan at Oakridge National Laboratory and 16.3 for Sequoia at Livermore National Laboratory), allowing simulations involving trillions of particles in cosmology as well as simulations of the electrophysiology of the whole human heart. The next-generation supercomputers are expected to reach the exascale (thousand times faster than the petascale) in the next ten years. There are technological challenges that need to be addressed to make this possible, such as dealing with the ever-increasing power consumption and the rise in the probability of component failures. The biggest challenges may come, however, from the scale of the data generated by the corresponding exascale simulations, which will require rethinking the way we perform simulations.

The traditional mode of operation for supercomputers is batch-mode. Executable code is submitted remotely to the computer as a "job" which the computer processes, in general, without interruption, before returning the result to the user. In this model, computing and data analysis are two disjoint, sequential processes. This separation, however, is already raising problems at the current petascale level in supercomputing and will be unsuitable for the exascale level. Exascale simulations are expected to generate petabytes of data per second. Transferring these data to other computers for analysis and visualization will become prohibitively expensive, slow, with serious problems of data integrity. In addition, most of the current algorithms for analyzing the results of simulations will not scale to datasets in the petabyte to exabyte ranges. Data of these magnitudes will also raise serious visualization challenges, as their complexity exceeds the capacity of the brain to extract information from images. Solutions to these challenges will lead to new approaches to supercomputing in which simulations will most likely be run interactively and analyzed *in situ*, i.e., with the analysis procedures and visualization routines running concurrently with the simulations.

Scientists at UC Davis are already actively engaged in developing novel approaches to simulations at multiple scales as well as novel methods for analyzing and visualizing their results. A wider integration of these efforts across campus would enable discovery in any field where simulations are important. Failing to support such integration would seriously hinder the possibilities for UC Davis to develop or participate in large-scale projects such as the Human Brain Project.

4. Big Data Exploration And Knowledge Discovery

Continuous technological advancements require continuous development in new approaches – theory and software for both visualization and modeling. Algorithms, software infrastructure, data analysis/information extraction tools, dimension and complexity reduction tools, visualization tools, as well as statistical tools for assessing the quality of the extracted information need to be developed hand in hand in order to result in effective methods that allow us to successfully tackle the Big Data problems that we are facing. Harnessing the combination of data analysts/ statisticians, developers of algorithms, developers of specialized systems (e.g., R) and programming language researchers is under way. We strive to establish a Data Science Institute that leads the nation in these efforts.

Data Analysis, Dimension Reduction, and Assessment of Knowledge Extraction: The UC Davis Department of Statistics is a modern department with a strong international reputation. Outstanding research on the analysis of complex data structures is already conducted there in collaboration with subject matter scientists. Collaborations include brain imaging (fMRI, DTI), spatio-temporal modeling, tracking and monitoring (behavioral continuous time monitoring for individuals of a population, continuous time traffic monitoring, gene expression data and time course for large number of genes and subjects, online monitoring), the analysis of time varying social networks, cosmology (cosmic microwave background), etc. These research efforts need to be promoted further, and their existence on the UC Davis campus also provides an excellent foundation for attracting leadership and expertise in the intersection of Statistics and Computer Science that is relevant to the Data Science endeavor. Relevant fields include statistical machine learning/data mining, statistical computing and computational statistics.

New methodologies are needed for assessing whether features (information) that appear to have been extracted from Big Data are features that are really present in the data or just noise (after all, much can be found in Big Data). The analysis of massive data sets often will make the calculation of approximate (rather than exact) results necessary, and these results are often stochastic in nature. Again statistical tools are needed to assess the quality of the approximations. Assessing the loss of information experienced by the application of necessary complexity and dimension reduction techniques is yet another important task when dealing with Big Data. The development of effective tools to tackle these questions requires a deep knowledge of the nature of the problem and the scientific goals, again necessitating the close interaction of statisticians with other data scientists, computer scientists and domain experts.

It is of highest relevance to note that without the availability of such tools data science techniques will not result in the significant research progress suggested by the growth in available data. In other words, the growth in data will not lead to the potential growth in insights without significant changes into how we do science and data analysis.

While some relevant expertise is already present on our campus, it is crucial for the success of the data science endeavor to strengthen this area further and to develop leadership and a critical mass in order to facilitate the amalgamation process outlined above. Equally crucial is to establish a culture and organization on our campus that supports and reinforces these efforts.

Algorithm development: The development of complex statistical methodologies for the analysis of Big Data and the development of novel algorithms needed for their analysis (e.g., parallel and out-of-memory algorithms) also need to go hand in hand. Understanding approximations and bounds to allow for approximate results is of immense importance for the analysis of Big Data.

Facilitating a closer connection of statisticians/data mining experts and algorithm developments on our campus, thereby enabling cross-fertilized research, will be another huge opportunity that comes with the data science Institute.

Software Infrastructure Development: Domain Specific Languages (DSLs). Developing new programming language(s) for data analysis is another opportunity. Rather than using general purpose languages, statisticians have had much success with languages and systems such as S/R, MATLAB and SAS. S/R has illustrated the value of using high-level languages tailored to the needs and uses of a particular science activity/set of users/community. We have to go further and develop more domain-specific languages that allow us to express the nature of the computations we want to perform but not how. This makes them more adaptable to different platforms and captures much more information. The challenge is to allow higher-level specification with improved performance, typically a trade-off. Now we can do both. However, there are other domains in which specialized languages would be valuable - e.g., specifying sampling schemes; describing data partitioning for distribution across file systems, clusters, CPUs and GPUs; describing updating algorithms/techniques; an extended graphical models language. The result must be scalable to the exascale.

Data Technologies for Science: A Data Science Institute provides the opportunity to also build new software infrastructure that industry is not likely to develop and that is applicable to the different sciences. Research in this field is needed since there is no single, general-purpose approach that is best for all computing. Big Data has already given rise to new technologies such as MapReduce, Hadoop, NoSQL databases and so on. These have typically been developed within industry rather than universities or "traditional research.” Just as general purpose GPU computing uses the hardware developed for video games to speed up large scale simulations, we can use the industrial tools that have been developed for commercial purposes to do scientific work. MapReduce/Hadoop is the most obvious example, but Google Earth and others are also important. The Institute will give us an entry point through which we can connect our research problems to industrial tools. The Institute will also provide the opportunity to enable researchers from different fields to develop infrastructure by facilitating collaboration across the disciplines and utilize the commonalities, benefits and efforts across disciplines.

There are also interesting opportunities for integrating data analysis and data storage and access tools. For example, as we develop new languages and databases, we can explicitly address how to integrate them, which has not been done in a widespread manner. Working with existing systems, we can develop approaches to allow analysts and scientists to develop code that runs within the database.

Data analysis and visualization technology: Making sense of the enormous volumes of data requires automated or semi-automated analysis using techniques that can detect patterns and trends, identify and classify anomalies, and extract knowledge. Visualization has been shown to be effective in giving an overview of large data, capturing intricate interactions between objects of interest, showing their evolution, and directing the data exploration and analysis tasks. Visualization is a crucial technology that can drastically enhance our ability to reason and manage large, dynamic data. Visualization is the most intuitive and convenient interface to navigation and manipulation of the data. UC Davis already has world-class visualization research programs, giving us a distinct competitive edge over other institutions. We need to develop new tools and theories in visual and statistical inference and learning for knowledge discovery from massive, complex, and dynamic data sets.

Most importantly, Big Data means “constantly changing”: what we will be doing in the short-term will change. We need to be agile and adaptable. As the nature of the problems, the data structures, software, hardware, etc. continue to change, we must be able to lead by being able to develop and try novel ideas.

5. Data Governance and the Social Aspects of Big Data

Several of the challenges and opportunities of Big Data at UC Davis go beyond the technical aspects of algorithmic innovation and hardware infrastructure and relate to social, legal, business, and ethical issues. As it becomes easier to generate, accumulate, aggregate, analyze, and share data, new complexities surrounding data emerge. Even when researchers have unconstrained legal access to their own data, questions of data security, transferability, and intellectual property become increasingly complex as the research scope expands and the number of stakeholders increases. Such concerns will only ramify over time, insofar as methods for producing, storing, sharing, searching, analyzing, visualizing, publishing, and profiting from research data will become ever more sophisticated, automated, and potentially lucrative.

These questions are especially pertinent for data that involves human subjects, where privacy, regulatory landscapes, and responsible innovation come into play. While standards have been established for human subjects research at various scales — from local ethnographies to international clinical trials — and while Internal Review Boards are generally equipped to evaluate research protocols at these scales, new methods for studying Big Data introduce unprecedented capacities to query, calculate, and extrapolate human subjects data in ways that may exceed the ethical or legal parameters appropriate to earlier modes of research. Research is now underway in the fields of medical anthropology, media studies, and Science & Technology Studies (STS) on the effects of these changes in different areas of society.

New challenges likewise emerge in the humanities relative to Big Data, particularly when the volume and accessibility of archival data surpass existing legal protections for research on cultural materials. For example, while the “Fair Use” doctrine of U.S. copyright law has long facilitated scholarly research on discrete quantities of print-based media, changes to U.S. copyright law under the Digital Millennium Copyright Act created obstacles for studying digitized materials that are especially salient when dealing with large online repositories of texts or other media. Even when such large databases and the computational tools for studying them already exist, humanities scholars are not always legally or financially able to search, analyze, or reproduce the data in all the ways they might hope to.

"Data governance" is the system of decision rights and responsibilities that describes who can take which actions with which data, when, under which circumstances, and using which methods. It addresses the issues described above as well as strategies for data quality control and management in the context of an organization. It includes the processes that ensure important data are formally managed throughout an organization, including business processes and risk management. Data governance aims to ensure that data can be trusted and that people are made accountable for actions affecting the data. At the same time, new media systems (e.g., Facebook) often disrupt existing solutions, and new data practices create the need to empirically study user behaviors, desires, and networks.

Where does the responsibility of data curation lie? Universities are being asked to curate data generated by federally funded projects. Increasingly funding is contingent on a long term curation and data access plan. There are economies of scale here which need to be explored.

Developing tactical solutions as well as institutional strategies to manage the data governance challenges of Big Data will be crucial to the success of the UC Davis initiative, and so it must incorporate research programs on the social, legal, business, and ethical aspects of Big Data in addition to the scientific and technical aspects. Understanding data governance and having mechanisms in place to deal with it as it changes over time will be essential for researchers working at new scales and using new tools to improve our ability to turn data into knowledge.

6. Big Data: Accelerating Innovation at the Interface between Disciplines

Most researchers do not view Big Data as an end in itself but rather as a means to an end. Big Data can both help us pose new interesting questions and assist us in pursuing elusive, impactful answers. A common feature of Big Data research is that it often strives to reveal new insight by identifying and explaining dependencies or complexity in previously unknown parameter interrelationships, extracted from massive and often structurally complex datasets.

Fascinating insights have already resulted from Big Data exploration in diverse domains such as genomics, climate change research, cosmology, and macroeconomics. We have provided a number of 'research vignettes' providing domain specific examples in appendices to this report. However, it appears highly likely that major breakthroughs will also occur at the intersection of disciplines. Combining sets of Big Data from different sources opens up a plethora of intriguing possibilities:

> Can a computer algorithm accurately predict individual lifespan, future health or perhaps accelerate treatment through evidence-based medicine? Integrating clinical, scientific and behavioral/demographic data is highly likely to identify new important correlations between lifestyle and health, potentially realizing the concept of personalized healthcare.
>
> As our global population continues to grow, high performance computing and Big Data approaches may play an important role in optimizing food production. By combining our knowledge of energy, water, seed, plant and soil biology, climate modeling, and social factors in new ways, we may enable a new field of data-driven agriculture to help the world feed its people more efficiently.
>
> Big Data techniques can help us identify underlying patterns that may predict the emergence of fruitful exploration between new combinations of particular disciplines or fields, or help us

more rapidly translate key findings in one field into another. Big Data analytics may also help us accelerate the translational time from basic research to technology development by enabling deeper analysis of past successes. Can correlations between scientific publications, or patents in one sectoral or perhaps geographic domain predict future value/wealth creation?

Real or near-real time information delivery is one of the defining characteristics of Big Data analytics; perhaps real-time stream computing can augment our senses by rapidly integrating disparate data sets to help us more effectively deal with life threatening natural disasters, avert impending financial market collapses, or alert us to other extremely dynamic time-sensitive events.

IV. TEACHING

UC Davis has taken some steps to ensure that all students can reason quantitatively. Our campus has made quantitative literacy a critical component of Core Literacies; as such, all students are trained to “reason quantitatively and to evaluate quantitative arguments encountered in everyday’s life.”[7] Courses have been developed to allow the students to fulfill their quantitative literacy requirements, mostly offered by the Statistics, Mathematics and Computer Science Departments, although some specific courses have been developed in the domain departments (such as Economics, Biological Sciences,…). A significant component of these courses is to train students to understand the relationships between events occurring in nature, data collected to study these events, and the implications of the analysis of these data for the understanding of these events. The democratization of Big Data, however, and the sheer size of the challenges they come with raise these requirements to a completely different level that cannot be addressed solely with the existing education programs. Big Data are pervasive and affect all aspects of our professional and social life; it is therefore essential that our community, including undergraduate and graduate students, staff, and faculty, be properly trained on how scientific, economic, or social values can be extracted from these data.

The Data Science Institute brings together a diverse interdisciplinary team of researchers from core disciplines in data science and domain sciences. This concentration of diverse talents and expertise provides a unique opportunity to create new training programs in data analytics, aimed at producing a new type of professionals and researchers capable of meeting the emerging Big Data challenges. These programs will be intended to establish new models for undergraduate and graduate education and training that transcend traditional disciplinary boundaries, and to engage students in understanding the processes by which data is translated to knowledge and understanding. They will directly benefit the undergraduate and graduate students that are involved with Big Data. It will also be the mission of the DSI to provide outreach programs to other groups on and off campus, and offer training in data science for the community at large.

1. Big Data Analytics Training for Postdoctoral Fellows, Staff, and Faculty

Colleges and core campus programs already offer workshops, lecture series and boot camps to support the ongoing professional development of their staff and researchers. The same vehicles should be used to promote Big Data analytics to staff and faculty. The role of these educational programs will be to promote understanding of Big Data analysis at a conceptual level, including methodologies for data handling, programming concepts, algorithmic developments, and statistical learning methodologies, as well as at a more practical level by offering short applied courses that will develop hands-on skills. Such courses will cover the use of specialized software tools for data analytics, the use of statistics techniques to analyze data, familiarity with cloud computing, analysis of risk, security and privacy concerns, and many other topics.

The structure of the DSI itself will promote education for staff and faculty. The DSI is designed to stimulate collaboration between core data science disciplines and domain sciences. It will reach this goal by providing space for interaction, both for permanent members of the Institute and for affiliate members through the concept of “hotel space.” The hotel space provides an opportunity for researchers who are not familiar with Big Data to collaborate with experts in data science in the early stages of their projects. This opportunity to collaborate will broaden and enrich their ability to approach Big Data challenges.

[7] Regulation 523 of the UC Davis Requirements for Higher Degrees

2. Data Science at the Ph.D. Level

The current structure of graduate-level education at UC Davis will facilitate developing and offering a graduate program in Data Science. Graduate programs are organized as interdisciplinary graduate groups, giving students freedom to transcend disciplines and areas of research. Students in certain graduate groups may also participate in a Designated Emphasis, a specialization that may provide a new method of inquiry or an important field of application. The graduate group and designated emphasis models would be well suited to directly support the educational charge of the DSI at the graduate level. We therefore propose to:

- Develop a Ph.D. program in Data Science (graduate group level). Students joining this program will have the opportunity to work directly with a faculty from the DSI, either from the core disciplines of Data Science, or from a domain science. These students will ultimately work in collaboration with researchers from both backgrounds, which will broaden their formation.
- Develop a Designated Emphasis in Data Science. Students joining this program are expected to be enrolled in a Ph.D. program in a domain science and demonstrate interest in developing a specialization in Big Data analysis.

3. Data Science at the Master's Level

As Big Data become an integral part of our life and a mixed blessing for industry due to the combination of opportunities and challenges they bring, there is a need to train a broader range of experts in Data Science that may not require a Ph.D. For example, expertise in Hadoop, the open-source framework for handling large quantities of distributed data, is in huge demand by many companies that cannot afford to train their own employees. There is a unique opportunity to offer Master's programs and professional certificates that will satisfy these needs while also generating revenue for the University.

The faculty members of the Data Science Institute will work in partnership with the appropriate departments to develop a core curriculum for Master's programs in Data Science. These programs will have a strong revenue-sharing component to provide incentives to the participating departments; alternatively, professional Master's programs might be developed. The structure will be worked out with the appropriate departments, including Computer Science, Statistics, the Graduate School of Management, Applied Mathematics, and others. The sharing of resources and the involvement of different departments will result in programs that can provide unique multi-disciplinary training opportunities.

One possible scenario is to offer two-year degrees with core courses in several departments, a suite of electives, and a three-to-six-month thesis on a project with an outside company. The courses will be shared between departments and will thus make it possible to offer curricula covering three (or more) disciplines -- the domain areas (Marketing, Healthcare), statistics, and computer science. These 'boutique' degrees would feature the following key elements:

- Multi-discipline training;
- Our Northern California location and proximity to Silicon Valley; and
- A thesis option to facilitate not only high-quality education, but also top placement with tech companies and promising start-ups.

A bold version of this initiative entails securing guaranteed placement of UC Davis graduates with companies facing Big Data issues. To be realized, this goal requires expertise for DSI faculty members to develop innovative course curricula meeting the highest standards, as well as vision and commitment from senior leadership both in the pre-launch phase and in the start-up mode. By selecting the finest students, arming them with cutting-edge knowledge and helping them tackle real challenges, we will develop the next generation of leaders whose success in the workplace will build the reputation of the degree program, the Data Science Institute, and the campus as a whole.

4. Data Science at the Undergraduate Level

DSI faculty members will have a leadership role in developing and delivering a GE-type undergraduate curriculum in Data Science. The program will build on the current Quantitative Literacy requirements; it will also provide students a general critical awareness of issues involving Big Data, and give them a minimum set of skills and methodologies for analyzing Big Data. This will prepare UC Davis undergraduates for the job market and will give them an edge when applying for graduate school in virtually any discipline.

Ultimately, we envision developing an interdisciplinary undergraduate program in Data Science to meet student demand for in-depth knowledge of the concepts behind Data Analytics for Big Data that goes beyond what they can get within a single department. There is no real precedent on campus for creating and managing such a broad undergraduate program; we therefore seek advice from the academic council on how to best implement this idea.

5. Outreach

The DSI will foster an inter-disciplinary dialogue on social, cultural and scientific issues revolving around Big Data by organizing an ongoing series of seminars. These activities are of crucial importance and will be one component of the DSI's outreach program with the greater community, both on and off campus.

6. Costs to Implement and Run the DSI's Education Program

The Data Science Institute's ongoing budget must account for the cost of developing and running these educational and outreach programs, including buy-outs from home departments for faculty members, TA support, and teaching equipment. While buy-outs from home departments will initially support the launch of the educational programs, new FTEs will have to be allocated to ensure continued cutting-edge training and research, to establish the graduate group, and to develop and deliver undergraduate and graduate courses.

In addition to general support staff, the revenue-generating Master's programs will also need the expertise of placement officers who will develop and nurture contacts with private industry partners and provide critical feedback to the Institute's faculty.

V. INFRASTRUCTURE

While agencies, national laboratories, and other national and state programs have already begun to invest in the physical and intellectual infrastructure required for critical breakthroughs in dealing with the challenges raised by Big Data, universities have also begun making substantial investments in new capabilities. Significantly, such capabilities take the form of new, sustained -- *not* one-time -- investment in three areas: faculty to envision, frame, and lead in the creation of new knowledge both directly in relation to Big Data and its application across the disciplines; technical staff, post-docs, graduate and undergraduate students to support the broad campus community in its movement to Big Data and to participate in knowledge creation; and key technologies both on campus and in partner institutions. Here we discuss how UC Davis can rise to these challenges with respect to the latter, i.e., IT infrastructure.

Central to the success of an interdisciplinary approach to solving the Big Data challenges is the necessity storage, computing, and network infrastructure to meet the needs of researchers to allow them to focus on the science, rather than the technological limitations preventing them from conducting innovative research. This infrastructure needs to provide the glue that enables access to technology resources, and manages those resources. It consists of technology, governance, and people. Governance describes *how* the resources – both technological and human – are controlled, accessed, and used, and by whom; technology implements the policies arising as part of the governance; and people manage and use the technology resources. The University Big Data infrastructure must go beyond today's approach of relatively discrete high performance computing clusters, large data storage arrays, virtual server farms, and fast, but separate research and core campus networks based in disparate locations on campus. The Big Data infrastructure must integrate these technologies, providing seamless access to compute, storage, and network needs that is scalable to accommodate varying research needs. The infrastructure must also lower the barrier to entry by providing a service-oriented approach that allows researchers with relatively little knowledge of the various computing components of Big Data to use the service and its underlying infrastructure. Significant value will accrue to UC Davis if we make concerted, scalable investments (on-campus and with partners) to realize an infrastructure capable of handling Big Data.

Given the great breadth and depth of its research, UC Davis could be well positioned to become a key player in enabling Big Data science. We already have several strengths and experience building the foundations of a Big Data infrastructure, such as:

- The use of high performance computing clusters in some academic disciplines on campus such as agriculture and environmental sciences, astronomy and cosmology, genomics (both plant and animal), geology, and other areas;
- Some shared high performance computing resources providing both shared infrastructure and expertise for researchers;
- High speed networking from the campus core network to other academic institutions and data repositories;
- Programmed improvements to the campus research network that will provide massive bandwidth with limited firewalls for researchers to lower network barriers to large data transmission on and off campus;
- Experience with large storage arrays and virtual computing clusters primarily used on the production (non-research) network; and
- An advanced planning effort with the UC Davis Health System to construct a modern data center on campus for housing, advanced networking and support of research, clinical and administrative computing systems.

However, this existing technological infrastructure is currently distributed in ways that make sharing nearly impossible. More than that, the costs and complexity of augmenting the current infrastructure to the scales needed to solve problems over the next five to ten years are intractable. To meet the future needs of Big Data-driven research, UC Davis must shift from discrete infrastructure components used by some academic domains on campus to a coordinated infrastructure that is flexible yet service-oriented so researchers can "just use it." This is the major leap forward needed to provide a world-class Big Data infrastructure.

1. What Our Model Must Encompass

Any Big Data infrastructure must accommodate the needs of researchers to easily share and work with their respective repositories of data, both on and off campus. This sharing of data, however, needs to be secure and in some cases confidential (for example, with human subject data, the interdisciplinary correlation of data makes individual identification a greater possibility). To address the corresponding legal and ethical privacy concerns, key parts of the data will need to be anonymized, while other parts remain unchanged. Some research disciplines do not have this problem; however, almost all of them share challenges of data integration across formats and other technological and disciplinary differences.

As part of our vision to generate a seamless environment for sharing the (literally) billions of dollars of research data in our clinical, agricultural, sociological, genomic, and other data stores scattered around the campus, we must provide a way to bring that data together into an integrated environment that ensures the requisite levels of security, integrity, and privacy.

Early models of Big Data infrastructure created a large compute and storage pool in a single location, to maintain control over the environment and provide high bandwidth for the greatest performance. Large strides in network performance and reliability have led to the widespread adoption of public cloud computing services, which provide vast compute and storage infrastructure that is shared among disparate organizations. Many public and private institutions are now choosing to extend or transfer commodity storage and compute capabilities to these cloud providers in order to increase their ability to scale or focus less on building computing environments and more on the mission of their organizations. Public cloud computing must be considered for the UC Davis Big Data infrastructure, alongside academic, higher education or campus-based private cloud models, and hybrid models that combine private and public cloud infrastructure, but with a seamless experience to the researcher.

We first review the current state of our infrastructure. We then present critical components needed to support Big Data research.

2. Current State of the UC Davis IT infrastructure

Campus resources from computation to network to facilities are described briefly in Appendix C. While these resources highlight shared, campus resources, most investments are in systems within faculty research groups or in relatively modest shared clusters originally supporting start-up for new faculty. Health System Clinical facilities are very substantial, but are not designed for general access.

To develop an exascale Data Science Institute with minimal cost, UC Davis must build on the capabilities that it currently has. The current on-campus data center and UCDHS data center are minimally adequate for initial Institute pilot projects, such as a small, petascale data store integrated with existing HPC systems. The joint shared UC Davis Data Center under development is critical to the future of the Data Science Institute. The ability to bring state-of-the-art computational and storage engines, as well as to link to even larger off-campus facilities, depends on the campus investments in this shared facility.

Similarly, while the campus network backbone and outbound connections have kept up with other Research-1 institutions, the Data Science Institute will require high-performance networking connections of 10 Gbits and, in some cases, 100 Gbits to key core facilities such as the KeckCAVES and other visualization environments, centers such as the Genome Center with major instrumentation enabling analytics such as DNA sequencing, clusters for high performance computing (HPC), and other facilities. Current outbound connections include multiple 10 Gbit links. Over time, UC Davis must apply developments of next generation networking including software-defined networking (to provide flexibility to optimize network usage) and speeds of over 100 Gbit today. It must provide connections to major resources, including cloud and HPC services, instrumentation at remote locations (for example, telescopes), and so on.

New models are needed to determine the scale of HPC facilities associated with the Institute. Currently, most facilities are owned by individual groups or part of relatively modest shared resources. As we move to facilities matched with Big Data research, we need to ensure there are technical staff to provide a range of data support activities, including algorithmic optimization, effective architectural design and optimization (including the computing, visualization and networking components *in toto*), basic data management, sharing, integration, preservation, etc. These facilities and staff will enable the Data Science initiative described here as well as a range of other, simpler data management needs of faculty (e.g., secure, trustworthy data storage and management processes for even “small data” that is critical to faculty research).

3. Critical Infrastructure Components to Support Big Data Research

We outline seven major service areas, all of which are essential in supporting state-of-the-art Big Data research. It is impossible to say whether, for a particular application, 10 petabytes is enough storage, 1 exabyte will lead to a research breakthrough, 100 processor hours will lead to the required insight, or if 10,000 hours would accelerate research results. Equally certain is the fact that no campus today is prepared for the revolution envisioned by leading researchers and federal agencies. UC Davis specifically is hampered by an inconsistent infrastructure and common services upon which to build.

Our recommendations do not restrict any faculty member from building complementary services. We do recognize, however, the need for a critical baseline of services, including sophisticated intellectual resources, to support the broad range of data-intensive research applications extant at UC Davis. In fact, we expect that as faculty explore additional capabilities, the Institute must have mechanisms and incentives for integrating those into the core capabilities initially provided.

Finally, there is considerable strength at both the Sacramento and Davis campuses; any successful initiative must provide a “virtual” world that shrinks these campuses into a single, comprehensive Big Data infrastructure. Beyond that, it must ensure that powerful resources aligned with specific research areas nationally or globally, as well as cloud services provided at massive scales, are available to UC Davis researchers as if simply in the laboratory down the hall.

3.1. Data Center for Large-Scale Services

“Computational Facilities supporting storage, computation, and related services”

Current Data Center: Current facilities on campus provide priority access for researchers to 3,300 ft^2 of machine room space. UCDHS currently manages approximately 4,500 ft^2 of machine room space primarily for clinical and administrative systems in Sacramento. There are other smaller computer rooms distributed across campus where both research and administrative computing are operated.

Future State-of-the-Art Data Center: The future data center would be a single location to integrate all the Davis and Sacramento-based research in a single state-of-the-art facility. The proposed data center has been designed and presented to the Chancellor to serve as a campus-wide facility servicing the Davis campus and Sacramento Health System researchers. When funded, it is planned to be available beginning Fall 2015, with 30,000 ft^2 of machine room for research facilities, data storage, high-speed networking links, and staff workspace.

3.2 Storage Services

"Storage Services on-campus and from cloud providers, including both high-performance and low-cost storage at the scales well over 100 petabytes"

Current Storage Services:

On-Campus Research Storage: There is no central research data storage service at UC Davis today, such as an integrated site (for both storage and collaboration) like UC Berkeley's Research Hub[8]. There are large storage services for administrative computing needs, which are not offered widely to researchers at this time.

Cloud (Off-Campus) Research Storage: UC Davis has made some use of vendor (cloud) facilities at Amazon and Microsoft and has worked with Internet2 to identify some models for broad institutional access to very large data storage (both high performance and lower-cost, archival storage), but these services have not yet been widely offered.

Future Storage Services:

A Hybrid Storage Service Model:

The model for Big Data storage for the Institute includes a combination of on-campus research storage, storage at external data repositories, and cloud storage with vendors. This model is essential to provide the flexibility, lowest cost, preservation and collaboration capability that Big Data science will need.

On-Campus Research Storage:

While not a large-scale facility, UC Berkeley's Research Hub model has the potential to increase collaborative sharing, and offers campus-based storage for researchers. This is a good near-term research collaboration and storage model for the first phase of the Institute, linking with external resources such as UC3 (see below) and vendor cloud facilities. These would be housed in scalable hardware configurations in the Campus Data Center.

External Data Repositories: The UC-provided Merritt University of California Curation Center (UC3) resource[9] would be an integral part of the storage infrastructure providing UC Davis researchers with long-term preservation and curation of data. UC Davis researchers have begun using this facility; the Institute would explore effective cost models and local consulting and support to enhance the use of this facility. Similarly, the LSST project plans an exabyte data center at NCSA.

Cloud Storage: UC Davis is actively working with Internet2, UCOP, and other UC campuses to establish contracts for Amazon, Google and Microsoft cloud storage and compute services. The hybrid storage model includes cloud services as a way to provide additional storage and

[8] https://hub.berkeley.edu/about/ . Research Hub includes support for individual scholars, project teams, departments pooling resources, and large research projects requiring formal data management services.

[9] https://merritt.cdlib.org/docs/merritt_handout.pdf

collaboration options. Increasingly, researchers at UC Davis are using cloud services to share research data and to host common data sets across institutions. As cloud storage becomes less expensive over time and technologies and infrastructure for transferring very large amounts of data improve, an increasing proportion of data may shift to the cloud.

3.3. Computational Services

"General and specialized computational services, including hardware, software, and control systems"

Current Campus Computational Landscape:

The needed computational elements for the Data Science Institute are all in use today at UC Davis, albeit in a highly decentralized way across campus and used for a mix of research and administrative computing needs. These computational elements include:

- High Performance Computing clusters;
- Specialized HPC services, such as GPU services;
- Shared computing resources such as GENI and XSEDE;
- Virtualization Services for traditional computing locally (on campus); and
- Cloud computing services – on demand HPC and traditional compute services hosted by a vendor.

While these services are in place today, they are part of a cohesive set of services that meet the needs of researchers. In many cases for HPC and GPU services, the barrier to entry is high, requiring the purchase of systems rather than using shared HPC clusters. Virtualization services for traditional computing are not used by researchers at all and are only in place for administrative computing needs on campus today. Cloud services are used by some researchers who know how to seek them out, fund and utilize those services, but are not offered centrally through UC Davis-contracted vendors (e.g., Amazon and Microsoft).

Future Hybrid Compute Model:

A Hybrid Compute Service Model

A similar hybrid model of compute services must be developed to serve the needs of the Data Science Institute. HPC and traditional computing, on campus and cloud services must make up the offering so researchers have options to suit their needs for projects both large and small.

High Performance Computing

HPC clusters form the computational backbone of much Big Data research, and UC Davis has significant existing HPC equipment and expertise in many academic domains. The creation of a shared HPC service would significantly help and support the efforts of researchers currently without the funding and need for individual HPC clusters. A shared service will also be developed for specialized HPC clusters, such as GPUs.

Virtualization Traditional Computing Service

A virtual service allows a pool of traditional computing capabilities (i.e., separate servers) on shared hardware, lowering equipment costs. A virtualization service would be established as a valuable complement to HPC for researchers who have less compute intensive needs.

Cloud Computing Services

Cloud computing services hold much promise for researchers, and would be an integral part of a hybrid compute solution for Big Data science. Large cloud vendors such as Microsoft and Amazon have vast pools of compute resources that can be rapidly provisioned to almost instantly scale from low to massive compute power. When cyclical needs subside, that additional compute capacity can be released. Because machines are at a vendor, they can easily be accessed, passed on among multiple institutions as needed.

3.4. Advanced Data Management

"Services to support both researcher and institutional needs for data discovery, archival, and data integrity"

Some of the advanced data management services to support Big Data and other data management needs on campus are:

- Archiving of data (both NIH and NSF require plans for this, and given the large amounts of data involved this will clearly be a Big Data problem);
- Curation and preservation of data in a way it can be used and shared;
- Analytics;
- Database administration;
- Query tools; and
- Data provenance infrastructure.

While UC Davis does not have on-campus infrastructure to support advanced data management needs today, there are elements in place at the Office of the President, i.e., the California Digital Library's UC3 infrastructure that is mediated by the UC Davis Library. But we lack any sort of coherent data management infrastructure beyond those very basic services of data archiving, description and persistent identification for sharing, and simple long-term preservation.

3.5. Data Movement and Networking

"State-of-the-art networking between computational elements, to researchers, and to external partners"

Current Campus Data Movement and Networking Infrastructure:

Much of the needed networking elements needed to support Big Data science are in use today at UC Davis. Existing infrastructure elements include:

- A dedicated research network supporting 10 Gbps connections to research systems on campus and directly attached to the campus network border allowing high speed connections to regional, national and international networks. Funding has been secured that will make an additional 40 Gbps of bandwidth available to research systems in support of large data transmission on and off campus;
- Two high performance border routers, one on the Davis campus and one at the UCD Medical Center, each supporting multiple 10 Gbps connections to CENIC, Internet2, the National Lambda Rail, ESNet and other regional, national and international networks;
- A high speed optical network that provides multiple 10 Gbps connections between the main Davis campus and the UCD Medical Center in Sacramento;
- A robust and extensive fiber optic cable plant on the Davis campus that can be extended into research spaces allowing for the provision of high speed network connections to the campus research and production networks; and
- A modern campus production network providing 1 Gbps and 10 Gbps connections into office and research spaces.

Future Data Movement and Networking Infrastructure:

Some of the data movement and networking services and enhancements that will be needed to support the data science enterprise on campus will include:

- Robust networking for collaborators within a building; the problem is that much of the horizontal wiring on campus is outdated and too slow for Gbit connections. Communication Resources' plans for upgrading and improving the horizontal wiring infrastructure should be advanced;
- Advanced networking to specific core facilities on both the Davis and Sacramento campuses will be needed to provide high speed networking between computing, visualization, storage and other critical resources supporting the Data Sciences Institute. This will largely entail the extension of fiber optic cables into research and collaboration spaces, and in the long term, will include upgraded electronics/optics to provide 100 Gbps connections to specific facilities and resources;
- High-performance networking to external cloud resources such as Amazon's EC2 service and access to national laboratories and data repositories; and
- Sustained support for staff and instrumentation to provide network and data transfer performance optimization services and support for software defined networking research.

3.6. Visualization/Graphics

"Software and facilities supporting insight through visualization"

The KeckCAVES on campus is an example of this type of facility. Other specialized visualization services, both on and off campus, will require infrastructure support such as high-speed networking. In fact, baseline data exploration/visualization tools and the expertise to apply them to data is a core infrastructure service in some disciplines (e.g., bioinformatics).

3.7. Expertise/Intellectual Resources

"Staff to provide brain trust of expertise in effective utilization of large-data resources in support of state-of-the-art Big Data research"

Because aspects of working with Big Data require techniques and tools that differ from those used for smaller data sets, researchers new to the field will require assistance. This means that the Institute must provide staff as a key resource for users. Among other areas of expertise such as system administration and network management, staff members must be available to assist with:

- Algorithm development and optimization;
- Tools and other software suitable for Big Data;
- Database optimization;
- Infrastructure optimization, specifically when designing infrastructure and tools to move data among different networks, servers, databases, and other storage resources;
- Domain-specific applications expertise; in many cases, developing this will require collaboration among the researchers and their students and staff, and the staff of the Institute;
- Specialized services such as anonymizing data, helping design secure architectures for managing and protecting data, helping researchers use these secure services;
- Governance-related expertise, for example to help develop policies suitable for the specific application or data set; and
- Connection to the Big Data resources of the Institute; while this may seem trivial, new faculty or faculty who have never used the resources of the Institute will undoubtedly find this complex and welcome staff to help.

People knowledgeable in the semantics of different fields are critical to interdisciplinary collaboration. These people will enable researchers in different disciplines to connect with one another through a common vocabulary.

4. Implementation

Here, we focus on the short term (meaning within 2 years of the Institute being approved). Our suggestions are driven by making the infrastructure as real as possible as the Institute is being implemented, so that faculty and other researchers can begin using it immediately, and, perhaps more importantly, accept that it will provide needed resources and assistance in the near term rather than see it as an abstract entity that may meet their needs someday. Our plan proposes to enable faculty to give feedback on the proposed resources very quickly.

Specifically, we recommend that the support staff of the Institute be built up quickly. This will give faculty confidence that UC Davis is going to be ready as everything goes on-line, and as people work with faculty, they will know what will be available and ready to use. Also, if faculty needs are not met by what is planned, they will be able to communicate their needs as the Institute is created. The new Institute support staff will be able to interact with existing staff in other departments to clarify infrastructure, staff, and training needs.

4.1. Administrative Issues

The first step is to develop a decision structure, which includes an oversight committee and rules of governance.

The next step is to develop a funding model and mechanisms to enable faculty to use the resources of the Institute. It is not sufficient to simply say that faculty who want to use these resources need to figure out how to pay for them, because many funding agencies consider those resources as overhead and will balk at funding their use as a line item in a grant proposal budget. For these researchers, one option to consider is a system of seed funding (either through money or through direct provision of resources, including staff time) to enable them to use the resources. For those whose grants and contracts do allow direct budgeting for the use of such resources, a rate model would allow the users to know what their charges will be, and plan accordingly.

In the longer term, administrative support will be provided for faculty and researchers to develop grants taking advantage of the resources of the Data Science Institute to study problems of interest to their disciplines. Simple access to a range of professional expertise will greatly ease the burden on faculty who wish to use the resources of the Institute.

4.2. Resources

The first step is to identify key initial resources, such as a Davis Research Hub (see above) linked with UC-wide UC3 facilities.

Complementing institutional or consortial cloud services, we will aggressively explore the role of commercial cloud providers (such as Microsoft or Amazon) as a seamless part of Institute resources, especially if such facilities allow us early access to substantial capability, while local cloud infrastructure is built up.

In the following section, we provide an at-a-glance summary of the key components of the Data Science Institute, and we propose both short- and longer-term milestones by which those components will need to be implemented so the vision for Big Data at UC Davis is realized.

VI. DATA SCIENCE INSTITUTE: PROPOSED IMPLEMENTATION TIMELINE

	0-3 months	3-6 months	6-9 months	12-24 months	30-36 months	36-48 months	48-60 months
INSTITUTE LEADERSHIP	• Begin **Director** search • Identify initial core from existing **faculty**		• **Director** hired • Launch hire of new **faculty**	• DSI **faculty** in place		• Growth through **grants & industry** partnerships	• Expand **outreach** programs
INTELLECTUAL RESOURCES	• Hire **core post-docs/grad students & staff** (Research/ algorithmic dev., IT & library services)		• Launch plan to hire broader DSI **staff** and **post-docs**	• Full DSI **staff** in place	• Implement **shared solutions & tools**	• Support DSI programs & services	• Support DSI programs & services
TEACHING	• Begin **outreach & awareness** campaign for faculty & students	• Launch **seminars**	• Launch **boot camp**	• Start **Master's, undergrad** programs			• Expand **educational** programs
SPACE	• Secure **core space** for DSI faculty and staff	• Establish **'hotel' space** for DSI affiliates	• **Education/training space** available • Begin planning DSI **building**	• Major collaboration for **educational & research** programs underway		• Expand **educational & research** programs, building on Consolidation Phase infrastructure	• DSI **building** complete • Expand collaboration program
TECHNOLOGY	• Begin design of **Basic Phase**	• Implement Basic Phase	• Begin planning for **Consolidation Phase**	• Consolidation Phase infrastructure in place • Begin **Expansion Phase**	• New campus data center complete • Implement Expansion Phase infrastructure		

2013 → 2018

Big Data Core Technology Infrastructure: Three proposed phases -- One initial (Basic), guided per the charge from the provost by a group of faculty, the Big Data Committee Chairs, and technology leaders convened by CIO Siegel; one guided by the DSI Director (Consolidation); and one as a state-of-the-art facility (Expansion), as the proposed campus data center is completed. A brief overview of major goals for each phase follows.

Basic Phase: Provide critical initial large-data services to address the needs of the broad UC Davis community, incl. the social sciences and humanities, while supporting researchers who need access to campus and cloud services now.

- Provide basic big data facilities that enable inter-disciplinary collaboration on and sharing of big data (e.g., UC Berkeley's Research Hub);
- Begin design of cloud services to be hosted at UC Davis; and
- Provide support for use of public cloud services (e.g., Microsoft Azure, Amazon Glacier, Google Cloud Platform) to enable current big data users to expand their research programs as needed.

Consolidation Phase: Consolidate and expand services to support the big data community and its initiatives (incl. health-related data, extending high-speed networking to research spaces, and greater availability of expertise); a key goal is to enable sharing.

- Draw upon experiences of big data researchers from UC Davis and other institutions;
- Apply lessons from current high-performance big-data facilities (e.g., SDSC or UW), and next-generation cloud services;
- Expand support for faculty, students, and other researchers; and
- Develop new services (e.g., advanced discovery, metadata, and secure access) incrementally.

Expansion Phase: Launch a fully-scalable facility that seamlessly integrates a consolidated service at UC Davis with off-campus cloud services.

- Includes metadata services, advanced search capabilities, and tight integration with HPC services, both on and off campus; and
- Build on the successful interdisciplinary services from earlier phases, with an emphasis on the ability to scale equally for performance, dissemination and outreach, and curation/archiving.

VII. APPENDIX A -- VIGNETTES: BIG DATA @ UC DAVIS

1. Big Data and the Humanities

In recent years, new possibilities for creating, accessing, navigating, and analyzing enormous online archives of cultural materials have triggered a surge of innovation in humanities research, particularly in the field of the "digital humanities." Currently, humanities scholars have legal or quasi-legal access to several large collections of digitized cultural materials, such as the Google Books collection and various image and video libraries around the world. While more and more cultural materials are coming online all the time—indeed, the entire Internet and its global history is now a major topic of humanistic study—the ability to integrate sophisticated analytic tools across diverse big datasets, correlating various kinds of relevant cultural, historical, sociological, anthropological, or linguistic information, is currently severely limited. Cutting-edge humanities research (e.g., at CalIT2) is already dependent on the infrastructure, accumulation, aggregation, and navigation of enormous databases (which will continue to grow all the time, insofar as cultural research requires the retention, characterization, and persistence of historical data in its totality). Humanities researchers have started to develop new computational tools for analyzing vast quantities of texts, images, audio and video recordings, interactive media, 3D object maps, demographic and economic data, and so forth: in effect, the entirety of recorded human history (or at least, whatever portion has been successfully digitized at any given point in time).

For significant and rapid innovation in next-gen humanities research, it will be crucial to develop data governance solutions to ensure legal and programmatic access to relevant databases and archives. Software and platform development for mining various types of content, particularly for discovering contextual patterns or evolving semantic structures in large archives of video files, audio files, interactive media, online communities, and video games, will enable real breakthroughs in the humanities and humanistic social sciences. Optimally, the new tools and platforms developed for analyzing Big Data will be usable by humanists who are not already programming experts.

These developments present the possibility of an entirely new way of doing humanities ("Big Humanities"). Examples of first steps in this direction are Franco Morretti's work on tracking centuries of linguistic and narrative changes in the evolution of the novel. Timothy Lenoir has constructed and then used large-scale patent databases to to assess the efficacy of federal funding policies for facilitating innovation in bionanotechnology. Many STS scholars have studied how disciplines change over time using large database co-authorship and co-citation maps. Other digital humanists have been developing tools to visualize and navigate large amounts of cultural data. Art history, archaeology, and history of science have been creating three-dimensional navigable models to better understand social change, using technologies such as the KeckCAVES at UC Davis, but creating and storing this data remains an unsolved problem. Integration opportunities across the UC Davis campus will be critical. For example, innovations in reducing the "dimensionality" of data in astrophysics can be usefully adapted to video archives, and genomic and statistical research on semantic structuring of data will accelerate the study of large text and media databases.

UC Davis is well positioned to take a lead in the area of Big Humanities. Through the Digital Humanities Initiative at the Davis Humanities Institute, UC Davis has attracted three major center grants in Big Humanities, including funding the Humanities Innovation Lab, the IMMERSe network for video game immersion, and the Mellon Research Initiative in Digital Cultures. The Center for Science and Innovation Studies has been holding high-impact workshops for the last two years on evolving intersections of scientific research, legal practices, data governance, and economic environments, raising awareness of the need for better ways to understand cultures of innovation. The interdisciplinary culture of UC Davis also uniquely sustains many ongoing collaborations between humanities, scientists, and social scientists.

The major risks for big humanities databases are that because data is expected to be usable for decades and centuries, the obsolescence for media of any type (documents, video, algorithms) is a serious concern that needs to be addressed through data governance and preservation policies. Integrating Big Data solutions with the digital humanities on campus could make UC Davis a leader in this new field. The opportunity costs of not pursuing new modes of data preservation, storage, retrieval, and analysis include the curtailing of humanities research possibilities, as well as the threat of very expensive restoration or the loss of our cultural history. The long-term impact of pursuing Big Humanities at UC Davis is a real understanding of the magnitude of change across societies and generations.

2. Big Data in the Social Sciences

Increasing the capacity for social scientists to collect and analyze new data sources characterized by unprecedented breadth, depth and scale will transform our understanding of the social world. Although the emergence of Big Data computational research has been slower in the social sciences than in such fields as biology and physics, it is occurring in companies such as Google and LinkedIn, and in government agencies such as the U.S. National Security Agency. The lagging development of Big Data social science in university settings is explained by the unique complexities of social science data, the resources needed to overcome those complexities, and the relative dearth of resources for the social science in most universities in the U.S.

The Big Data used in computational social science vary widely in type, scale and complexity. Examples of existing research illustrate the diversity. Psychologists and neuroscientists code data from fMRI scans of human brains to investigate the interactions of neurons in large-scale neural systems and the cognitive mechanisms underlying human behavior. Political scientists mine data in extensive parliamentary and congressional speech archives, using natural language processing and textual analytic methods to extract insights about democratic representation. Economists create large-scale, multi-year micro-datasets by linking administrative records on retirement and disability benefit payments from the Social Security Administration, income records from the Internal Revenue Service, and health care services data from the Medicare and Medicaid programs to investigate the relationships between labor force participation, earnings and health. Sociologists use science journal and citation databases to examine the organization of academic science, the networking of individuals and institutions that produce innovations and the impact of the Internet on knowledge in society.

Big Data used by social scientists share the characteristics of Big Data used in the physical, natural and biological sciences – they are large in size, unstructured and multidimensional – but they have additional complexities that arise from their socially-generated nature. These data result from human interactions – with each other, with businesses and with social institutions – that are recorded via websites, mobile device, distributed sensors in the build environment, and the mundane completion of administrative forms. They are often collected by private companies and therefore proprietary, although “ownership” and “control” of such data is contested. These data are commonly considered “sensitive” since their distribution can have significant implications for the social, economic, and health outcomes of the individuals who are the original source of the data. Overcoming these complexities will catalyze the advancement of big-data social science that addresses issue related to the social world that are not directly linked to business profits or homeland security. Addressing the demands of big social data requires the concerted cultivation of institutional partnerships, computing infrastructure and interdisciplinary collaborations. It entails the development of dynamic data sharing partnerships between universities and the data-collecting businesses and government agencies, along with robust systems for secure distribution, use monitoring, encryption and anonymization of the data. Furthermore, to realize the potential of Big Data for the social sciences, interdisciplinary collaboration is essential: discovery requires that the application of advanced computational expertise be guided by social science understanding of human interaction and how to study it.

In such an academic environment, computational social science promises society a vastly enhanced understanding of individuals, collectives, and institutions, how they act and react in real social (and natural) environments, and what policies and practices may generate more positive outcomes. For example, we may identify the dimensions and interconnections of the multiple social networks that create a society, and how they evolve over time. Mobile phone usage data can be used to track the movement of people over time to understand how pathogens spread, or in real time to target the distribution of aid during crises that involve mass population displacement. By linking multiple administrative data source, we will identify the early childhood experiences that affect later educational and labor market outcomes.

3. Big Data in Cosmology

Inverse problems in science and engineering often are challenged by noisy, incomplete, and complex data. Often these problems are of high dimensionality. Usually it is the outliers that drive the instabilities in finding the solution. Statistical algorithms can offer solutions in some cases, but are not scalable to the Petascale; dimension reduction is required. One example from astrophysics is "photometric redshifts" -- estimating the redshift of galaxies (in data collections of billions of galaxies) by using the colors of the galaxies. The nature of the data makes it necessary to account for uncertainties and outliers as well as the impact of multiple $\chi 2$ minima. Innovative Bayesian statistical algorithms that utilize all the information to reduce the dimensionality are being developed at UCD. While the solution to this inverse problem will ultimately impact our knowledge of the nature of dark energy, any such novel algorithm will have application to similar challenges in all areas of science and engineering.

Cosmologists and high-energy physicists are excited about what they think may be a revolution in fundamental physics. Their dream is to understand the physics of "dark energy" – the mysterious late-time acceleration of the universe. Recognizing that cosmology is an observationally driven field, astrophysicists worldwide have put their hope in the Large Synoptic Survey Telescope (LSST). The cosmology group at UCD is heavily involved in making LSST a reality. The Astronomy and Astrophysics Decadal Survey "New Worlds and New Horizons in Astronomy and Astrophysics," recently convened by the National Research Council for the National Academy of Sciences, ranked the LSST as its top priority for the next large ground-based astronomical facility. Many regard LSST as the lighthouse project for looming Big Data challenges that are faced by most areas of science and engineering. Discovering the unexpected in petabytes of data is an exciting challenge with potential for significant spin-off.

LSST will generate several hundred Petabytes of data. Thirty terabytes of data will be produced nightly, and over one million alerts per night will be issued worldwide within one minute for objects that change in position or brightness. Mining these data quickly and efficiently for the known knowns and the unknown unknowns presents unprecedented opportunities as well as object classification algorithm challenges. Dedicated high performance computing facilities (up to 1.7 PFLOPS) will process the image data in near real time, with full-dataset reprocessing on annual scale. The nature, quality, and volume of LSST data will be unprecedented, so the data system design requires petascale storage, terascale computing, and gigascale communications. In order to develop new algorithms and database technologies, simulations of LSST observing (in collaboration with Google) are even now creating Petabytes of data per month, which have to be analyzed. Finally, the multiple cosmological simulations of the LSST survey (collaboration with LLNL) will require exascale storage and PFLOP compute capability to generate and analyze.

4. Big Data and Environmental Sciences: Opportunities

Human actions have transformed the world in a geological blink-of-an-eye. Rising levels of atmospheric CO_2 and climate change; heightened pressures on food production and security; clean air and water; land cover/land use conversion; the sustainability of natural ecosystems services;

development in regions vulnerable to natural hazards; biodiversity declines…these are just some of the changes that are threatening the sustainability of the Earth system. Yet unlike the disciplinary roots of science, these changes cannot be understood through the lens of single observations in space or time; nor are they amenable to single disciplines working in parallel or isolation. Rather, the environmental challenges we face will increasingly require an understanding of the very nature of complexity, encoded in huge sets of data and understood via interactions that play out at disciplinary interfaces – epitomizing what Big Data is all about – integration of complexity. Here, we highlight several emerging areas related to human/environment interactions that we believe are ripe for Big Data exploration and investment at UC Davis:

- UC Davis is uniquely suited to develop, apply and lead Big Data approaches in the area of eco-informatics, especially as it relates to the National Ecological Observation Network (NEON). This multi-million dollar program was recently conceived by NSF to advance our understanding of natural ecosystems, their services, and their responses to change within the US. NEON is a test-bed for Big Data; and the funds will continue to be invested over decades, including high frequency monitoring of ecosystems arrayed across different eco-regions. This program will be transformative; it's reasonable to expect at a "before NEON, after NEON" characterization of ecological assessment and monitoring. Investment in this arena of Big Data is needed to promote excellence in the ecological and environmental sciences at UC Davis, not to mention forge new synergies between natural scientists, mathematicians, computer scientists, and policy makers.

- The college of Agricultural and Environmental Sciences has been leading California, the US and the world in areas of food production, sustainability, security and risk – and Big Data investments are necessary to continue setting UC Davis apart. For example, Big Data will no doubt hold insights into the essential question – how can we feed a growing world population, given the pressures on resources, and environmental tradeoffs? Large sets of data on crop production, environmental conditions, economic models, environmental effects, nutrition and human decision making – from plot to regional and even global scales – makes this a growing Big Data topic in which UC Davis can lead the charge.

- Another Big Data challenge involves climate change – including science, impacts, mitigation and adaptation. At UC Davis, teams of researchers examining all aspects of climate change are in place to take advantage of Big Data. This area of Big Data emphasizes multi-dimensional stochastic modeling, regional climate downscaling, ecological, oceanographic, and geological processes, climate change through deep time, complexity and interactions; human behavior and decisions making; engineering based assessment of climate risk mitigation; and economic forecasting. Further, the emergence of disease in animals is highly correlated with climate change, posing serious risks to human health. Natural linkages between researchers in Vet med, genomics, ecological and the geological sciences are poised to explore these topics via Big Data integration at UC Davis.

- A fourth area involves the study of physical changes in Earth's landscape, associated with earthquakes, volcanoes, landslides, floods, and (the largest influence of all) human activity. The physical landscape on which we live is changing at an unprecedented rate, with corresponding costs to infrastructure and human life and safety. In order to document and understand this change, data are being collected through space-based remote sensing, land and ocean-based sensor networks, and airborne scans, by organizations as diverse as Google, NASA, NSF, NOAA, and even state and local communities. Complementing these efforts are large-scale models of the physical processes involved in landscape change, using high-performance computing, along with innovative methods for visualizing and interpreting multiscale, multidimensional data. UC Davis plays a leading role in this area through cross-cutting research in the physical sciences, earth sciences, engineering, computer science, and mathematics, and partnerships with national labs, government agencies, and national high performance computing centers such as XSEDE and NCAR-Wyoming-Yellowstone.

In each of these Big Data areas, several pressing needs exist – most importantly perhaps is the need to develop a campus-wide meeting-grounds, one that encourages creative, if not unorthodox approaches to Big Data at UC Davis. The cost of non-action can't be overstated; not only would the land grant institution be marginalized, UC Davis would fall behind competing institutions, and, most importantly, fail to train the next generation of thinkers and innovators in the area of coupled human-environment systems. Indeed, one can imagine a campus-wide Big Data Institute at UC Davis, whereby data-mining techniques developed in the area of genomics, leads to the next key innovation in sustainability science or discovery in climate change. In terms of technology, hyper- and multi-spectral satellite imagery, high fidelity and high frequency measurements of the atmosphere, geosphere, hydrosphere and biosphere, computer modeling across multiple dimensions of space and time, scenario forecasting into year 2100, and economical models that couple the human-environment system will shape computing demands. This means Giga to TerraFLOPS – and data storage capabilities that exceed hundreds of terabytes. NEON alone, for instance, will provide data-products that span point measures of such attributes as soil moisture (taken every second, from 18 different sites, for over 17 years) to multi-annual flight overpass of wLiDAR.

5. Big Data in Computer Science

Scalability - the design of systems that remain efficient even as they grow arbitrarily large - is the central concept in Computer Science. It brought us the exponential growth of computer memories, the internet and mobile networks, distributed data access, and on-demand search and ranking. Looking at our efforts through this lens is crucial for Big Data. Here we highlight a few key Computer Science technologies that seem critical.

First, scalability is getting harder. Moore's Law, that (transistor) density doubles every two years, has held for forty years. But recently it just means that hardware is getting smaller, not faster. Speed improvements come from parallelization - computing clusters, massively parallel graphics processing units (GPUs), and distributed services in the cloud. Some problems are "embarrassingly parallel," but many are not. While Davis boasts some great expertise in advanced hardware design, GPU programming, and distributed systems, parallelism is going to have to become a much more central topic.

Scalability is a function of data complexity as well as size. Our signature projects like the LSST and NEON include lots of image data; medical data does as well; and much of the interesting social science data is text. We call this "unstructured" data, but more accurately it is complexly structured. A surprising amount can be done by representing it as sparse vectors (e.g., "bag of words"). But knowing how verbs relate to nouns, how gravity structures galaxies, or that an animal has four legs helps specialized fields like computational linguistics and computer vision to go much farther, producing, in other domains, machines that "understand English" (e.g., Watson) or "see" (e.g., driverless cars). Unfortunately Davis has very little expertise in these areas. This should be a hiring goal.

On the bright side, we are world-class in computer security and cryptography; in networks, including sensor networks; in software engineering; and in visualization. We should plan to deploy these strengths, without detracting from the research, by adding allied staff to the Center. Database design (as distinct from data mining) enables fast access to structured data. Here, we could use both more faculty and support for deploying the expertise we have.

Scalability is the goal of Computer Science theory. We are strong in cryptography and computational biology, but require expertise in Big Data technologies like streaming and sub-linear algorithms, data sketches and core-sets, and basic graph algorithms. Theory is one area where academic research supports industrial research rather than trying to compete with it.

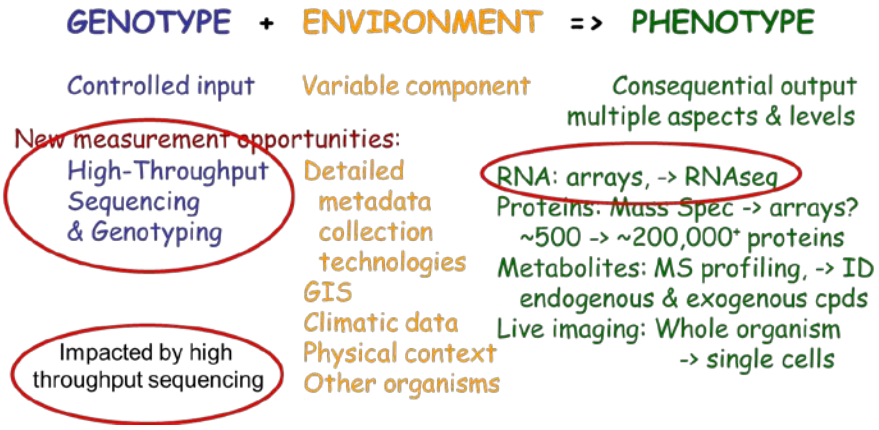
One of the challenges of Moore's Law is that, as computing systems shrink, the power density and heat load grows, along with greater complexity to remove that heat load. UC Davis' expertise in "green Energy" initiatives includes two "green computing" start-ups. SynapSense is a thriving Sacramento business minimizing energy use in cooling Big Data centers, and Ennetix optimizes network router energy use as a function of network traffic. Minimizing energy use should be a core goal in our infrastructure plans, as well as aligning with research relating to energy efficiency. A campus-wide Big Data initiative will accelerate our insights into realizable cost-savings for *exascale* computing systems and beyond; it also represents a Big Data problem in itself, one that will benefit from the multidisciplinary strengths of the campus, in Electrical Engineering, Computer Science, Materials Science, as well as in developing energy-related testbeds and startups.

6. Big Data in the Life Sciences: Opportunities and Challenges

UC Davis has probably the largest and most diverse biology faculty in the world. Multiple research groups are world leaders and are well placed to utilize the opportunities afforded by Big Data approaches. One of UCD's attributes is its collaborative mix of hypothesis-driven and more data-driven research programs. Worldwide, biological paradigms are being worked out in a limited number of genotypes in a few model species. UC Davis has a unique opportunity to determine how these paradigms hold true across biological diversity through comparative functional genomics. Our ability to achieve this will be in part determined by our ability to handle Big Data.

Biological research is moving from a data sparse to a data (over)rich reality. Disparate biological fields are moving at different rates in incorporating Big Data components. Ultimately, however, virtually every aspect of biological research will be impacted by Big Data. This is driven by technological advances that allow the acquisition of unprecedented amounts of data (Fig.). For example, a single DNA sequencing machine can now generate in less than four days as much sequence as was known and stored in 2009. Similarly, detailed GIS, satellite imagery and climatic data combined with comprehensive analyses of biological diversity is revolutionizing ecology. Also, imaging technologies from the single molecule and cellular levels to the whole organism and landscape levels are providing unparalleled resolution of life's processes.

A central principle in biological research is the interaction between genotype and environment that results in the observed phenotype. Multiple aspects of these components can now be measured with unprecedented granularity (Fig.), all of which generate vast amounts of data – much of which is unstructured. Data handling, analysis, and modeling rather than data generation are increasingly the rate limiting steps.



Very large amounts of sequence and other information are becoming available for many organisms. The reliance on model systems is diminished. High resolution and enormous datasets on genotypes and phenotypes are being generated at multiple levels: DNA, RNA, protein, metabolite, whole organism and microbiota. Biological phenomena need to be considered at the systems level as networks of interacting networks rather than isolated molecules or pathways. Such approaches require major data handling, analysis and modeling capabilities.

The social consequences of large amounts of biological data particularly that from commodity DNA sequencing is uncertain but will certainly be profound. Hundreds of thousands of humans will be sequenced and many attributes profiled in the near future. The genetic predispositions and molecular bases for many normal and pathological traits will be known. Big Data will result in the reversal of the knowledge paradigm; e.g., individuals will know about themselves than their doctors and insurance companies. This raises societal/ethical/legal issues of confidentiality and (mis)use. There will be opportunities for misinformation and misunderstandings and there will be a need for a legal framework to mitigate against abuse. It is essential that the University curriculum is adjusted to provide a genetic literacy to the next generation of medical practitioners and decision makers as well as the general public.
One of several biological strengths at UC Davis is plant, animal, and microbial genomics. The Genome Center provides enabling technologies for groups across campus. It generates terabytes of data per week and has data storage capacity approaching a petabyte. However, the imminent increases in amounts of data will overwhelm its current capacity and infrastructure. A major challenge is the integration of terabytes of data being generated locally with the exabytes of data being generated throughout the world, particularly by the BGI.

It is critical that the campus has the ability to acquire, curate, query, and distribute petabytes and ultimately exabytes of data of many different types. These data types include but are not limited to genomic, image, physiological, and biochemical data. We need to determine the balance between what should be performed locally and what should be outsourced to different locations and organizations. We clearly need both but the optimal balance is at present unclear.

Not developing and maintaining state-of-the-art capabilities to use Big Data (generate, distribute, process, analyze, and model) is not an option if UC Davis is to maintain its position as a leading biological research institution. Future funding research either to individual researchers or groups of faculty centered on our existing strengths will be predicated on our ability to handle Big Data. The ability to handle Big Data will also be a major factor in attracting and retaining excellent faculty in the biological sciences. Without significant strategic investment, UC Davis will be relegated to a second class research endeavor. Conversely, investment in building our Big Data capabilities will ensure our ascendency in the life sciences in both the immediate and longer term.

7. Big Data and Precision Medicine

Current advancements in genomic analysis have generated predictions that within five years, sequencing a person's entire genome will cost less than $1,000. This will have a dramatic effect on the routine use of whole genome sequencing in a new era of biomedical therapeutics often described as "precision medicine.” In such a setting, whole genomes of patients will be routinely sequenced and analyzed. The analysis will inform clinicians as to whether patients have certain mutations that make them more or less susceptible to specific treatments. This will lead to tailored therapeutics based on genomic data.

The data implications of precision medicine and routine whole genome analysis are significant. As noted by Haussler and colleagues in their white paper entitled "A Million Cancer Genome Warehouse,” to avoid piecemeal approaches with multiple whole genome sequences for the same individual, a natural evolution will be to have whole genomes sequenced once per person and warehoused for subsequent analysis as needed clinically or scientifically. In addition, we are likely to see individuals who have cancer have their cancers 'sequenced' as well. Using data from the California Cancer Registry (www.cancer-rates.info/ca) there were 153,235 cases of cancer in California for the latest year of available data (2009). If each patient had their personal genome and their cancer genome's sequenced, this would be ~306,470 sequenced genomes. If one considers that each whole genome sequence renders approximately 300Gigabytes of data, this would translate to 87.6 Petabytes of data **per year**. Once whole genome sequencing is routinely done as part of clinical care, this is a substantial under-estimation of the Big Data landscape in clinical medicine as it is restricted to only cancer patients.

The ability to analyze this genomic population dataset will undoubtedly lead to improved understanding of mechanisms of disease as well as therapeutic targets. However, the ability to efficiently store, manage, and analyze this amount of data will require new innovative Big Data approaches. Data reduction techniques along with more sophisticated analytics across populations that combine genomic and phenotypic data from electronic medical records are two key success factors in being a leader in this area.

UC Davis is well positioned to lead Big Data research in support of precision medicine. It has a Genome Center which is recognized for its high-quality genomic research and high-caliber scientists. It also has a Health System which recently achieved HIMSS Level 7 status, a national recognition of its advanced use of the electronic medical record. Such a level has only been achieved by ~2% of US hospitals to date. The combination of the Genome Center, Health System, and other engineering and analytics disciplines across the campus positions UC Davis very well in the area of data science and analytics in the era of precision medicine.

A major risk for the health sciences in not pursuing big data science will be to become significantly compromised in our institutional ability to compete effectively for research funding from the National Institutes of Health (NIH) and other sources of funding for biomedical research if we do not develop our Big Data potential. In addition, we would miss a significant opportunity to become one of the leading institutions in the world in leveraging Big Data in biomedical science. The future scientific discoveries will certainly require this infrastructure and being a leader in this area will attract world-class scientists, thus providing students unique learning and research opportunities.

APPENDIX B: INFRASTRUCTURE AND SERVICES SUPPORTING RESEARCH AT UC DAVIS

NETWORKING, HOSTING & STORAGE

DATA NETWORK

All UC Davis faculty have access to the campus high-speed data network, which includes:

- A **dedicated research network** supporting 10 Gbps connections to research systems on campus and directly attached to the campus border allowing the transfer of large data sets to regional, national and international partners;
- 2 border routers supporting **multiple 10 Gbps links to internal and external networks --** one on the Davis campus and one at the UC Davis Medical Center -- that provide high-speed connections to CENIC, Internet2, the National Lambda Rail, ESNet and other regional and national research networks;
- **High-speed optical network** that provides multiple 10 Gbps links between the main Davis campus and the UC Davis Medical Center in Sacramento;
- Three **10 Gbps connections to CENIC**, with a 4th planned for installation in early 2013; and
- Extension of the **campus fiber optic** backbone into research spaces allowing 10 Gbps connections to the campus research and production network.

Redundancy – The UC Davis network is architected to ensure that the failure of border functions at one campus will result in an automatic re-routing of traffic to the other campus border equipment.

ADVANCED RESEARCH NETWORK

In summer 2012, UC Davis was awarded an NSF grant of over $992K. The project is co-led by Computer Science Professor Matt Bishop and CIO Peter M. Siegel. It aims to improve the campus research network so research projects requiring high performance networking for moving large data sets or performing remote computations can take full advantage of the throughput of high-speed networks connected to the campus border. Planned improvements include:

- The installation of four additional 10 Gbps circuits to the research network
- An interface to the Global Environment for Network Innovations (GENI)
- Support for Software Defined Networking (SDN) research; and
- Staff and instrumentation to assist with network performance optimization.

DATA CENTER

Current facility and services:

- IET manages ~3,000 sq ft of machine room space at the campus data center and the Watershed Sciences facility.
- After virtualizing over 250 systems and migrating administrative systems to the Quest data center facility in Sacramento, space is now available at the data center to support HPC and other research needs.
- As part of the Quest migration project, funding is available to offset the cost of housing research systems in the campus data center.

Joint Davis-Health System initiative to replace the old data centers at both campuses:

- The chancellor has approved building a multi-phase, 70,000-sq. ft facility (30,000 sq ft of server room space, plus office space, mechanical rooms, etc.), plus 18,000 sq ft of undeveloped space adjacent to the data center for future expansion (if needed) on the Davis campus.
- The project team is reviewing options for commercial partners to help build and manage the data center, or do it within UC Davis. Completion of the construction phase is expected in 09/15.

HIGH-PERFORMANCE CLUSTERS

The current UC Davis data center provides co-location services for administrative and high-performance research computing, with 24x7 monitoring and operational support, for single server or full-rack needs per standards developed by the Server Room Space Evaluation Committee:

- **Racks --** Can be provided by the client or IET, and must be APC AR 3100 or AR 3107 (i.e., industry standard 19" racks). In many cases the most cost effective solution is to have the vendor provide their own rack, install it at their factory, then ship the completed rack over, using our rack and installing the equipment on site at UC Davis.
- **Power --** As many power circuits can be provided as needed. The researcher provides the power distribution units (PDUs) as part of the cluster specification (IET can help determine the best PDU to deliver power most efficiently). PDUs should be 208V 3ph 20A or 30A. We recommend APC AP8865 (IEC C13/C19) or APC AP7862 (5-20). These are

non-switched; switched versions are available.

- **Cooling and power density --** Campus facilities cool racks up to 15kw/rack. Higher density requires special consideration, may involve an additional cost, and may require housing at a remote facility (e.g., San Diego Super Computer Center).
- **Networking --** IET provides network drops to racks. The client is responsible for any networking within the rack. Generally clusters are configured with a single 1 or 10GB network connection to a client switch in the rack. See networking rates.
- **Housing rate --** $115/month. Costs are covered through an allocation from the provost for research computing in IET facilities effective FY1213. Researchers are responsible only for networking costs when research clusters are housed in IET managed facilities. See data center rates.

Assistance with the design and RFP process for research clusters (e.g., evaluation of processing capability, power, heat load): srsec@ucdavis.edu .

HPC SUPPORT PROGRAM

Under the management of Professor Louise Kellogg (Geology), the campus launched a high-performance computing facility in 2009, following a facility established by a completed initiative in Computational Science and Engineering. The current program is offered through a partnership with Provost Hexter, Vice Chancellor Lewin, Dean Ko and Campus CIO Siegel, and ongoing funding is under development. The facility currently manages 16 clusters with a total of 632 nodes (for CA&ES, UCDMC/health system, Engineering and MPS). One is a Tier 3 cluster for the Large Hadron Collider (with approx. 176TB of storage), and 2 or more clusters are under design, pending funding.

GENOME CENTER

The Bioinformatics Service Core provides expertise and infrastructure for the acquisition, curation, analysis, and distribution of large complex biological datasets as well as develops and performs computations, analyses and simulations addressing a wide variety of questions from genomics to network biology. The Core has seven staff members with overlapping expertise in computing infrastructure, Web/databases, scientific programming, biological annotation and statistics. The Core provides bioinformatics support for the wetlab service cores as well as for individual researchers with individual bioinformatics needs. The computing infrastructure of the Core includes three high performance computing clusters: 110- and 80-node CPU clusters and a GPU cluster with a total of 3,584 cores; data storage servers of 300TB; several ultra large memory machines with up to 512GB of RAM; a preserved instance in the Amazon Cloud and tools to access scalable resources provided by the Amazon Cloud Services. See http://bioinformatics.ucdavis.edu/ for more details and recharge rates.

VIRTUALIZATION SERVICES

Virtualization reduces overall costs and hardware needs by allowing applications from different campus clients (from departmental administrative systems to faculty-led state medical services) to share the same hardware.

CAMPUS IN-BUILDING WIRING

An initiative is underway to address the most critical building infrastructure ("horizontal") wiring needs on the Davis campus. It focuses on upgrading a number of buildings, prioritized by the deans, to provide reliable and high-speed connectivity to faculty research and classroom facilities. In August, Provost Hexter committed $2M in central funds for FY 12-13 and anticipates funding a total of $6M over the next 3 years. The colleges that decide to participate contribute to help support infrastructure improvements; central funds are allocated at a rate of $2 for each dollar of matching school or college funds. See list of buildings & program description .

AMAZON WEB SERVICES

UC Davis is increasingly using Amazon's suite of cloud computing and storage services to meet research, teaching, and administrative needs. These computing services make it possible to quickly create server and storage capacity in the cloud without having to buy hardware, software or space. Subscribers pay one monthly cost for only what is used. New Amazon services include: Redshift, a fast, petabyte-scale data warehouse service in the cloud; and Glacier, a low-cost, secure data storage and archiving solution.

WIRELESS

UC Davis operates a secure and encrypted wireless network that covers most major campus areas, including library spaces, classrooms, residence halls, research space, conference rooms, offices, and other collaborative areas. UC Davis is a member of *eduroam*, a service that allows users from universities worldwide to access the Davis network with their institution's credentials and gives our community the same privileges at other institutions.

IT SECURITY

Campus network security services include Virtual Private Network (provide secure connectivity from off-campus locations); firewall and host level security management services; and intrusion detection, prevention and security alert services. In addition, authentication passphrases meet NIST 800-63-1 specifications, and allow access to information resources held by federal government agencies and other academic institutions.

PRODUCTION SERVICES

SIMULATION AND MODELING SERVICES

2D and 3D animations, simulations and interactive services are available to support teaching, online content development, and scientific research. Common applications include Flash, After Effects, Maya and ZBrush.

DIGITAL IMAGING & PHOTOGRAPHY

Professional studio and on-site digital photography services can support scientific/ research documentation and teaching, and high-resolution digital imaging and 35mm scanning can capture archival work for digital materials.

ILLUSTRATION AND GRAPHIC DESIGN

Services include custom illustrations, brochure and poster design, scientific illustrations, cartoons and advanced PowerPoint development.

AUDIO/VIDEO

Services are available for recording and distribution of lecture series, seminars, interviews, lab demonstrations or field studies that can be delivered in a variety of formats (e.g., live/on-demand webcasts and audio podcasts), as well as professional post-production editing and development/distribution of on-line content.

ITUNES U

Many researchers turn to iTunesU to publish lectures, news stories, research findings, etc. Most of the content is available free and to anyone; some is restricted to the campus community, and is available only with a university password.

COLLABORATION AT A DISTANCE

VIDEO & WEB CONFERENCING

A 24-seat distance-learning classroom, a 15-seat videoconferencing room, and a portable videoconferencing unit are available for local, national and international single/multipoint connections. Web Conferencing uses Adobe Connect (audio/video chat in real-time, a shared white-board, screen sharing, and embedded Flash or streamed PowerPoint).

ONLINE COLLABORATION, WIKIS

SmartSite comes with a set of tools and the ability to create a virtual space to facilitate collaboration with colleagues from around the world. Confluence is a wiki maintained by IET for those who wish to use online writing, editing, and simple administrative tools for research or collaborative projects.

CONSULTING SERVICES

Professional consultants are available to assist with a range of research-related needs, including:

NETWORKING – Network performance engineering, application tuning and remote management.

COMPUTING HELP DESK – Technical support and consulting services at no cost on various software and campus applications.

IT PROFESSIONAL SERVICES – Specialized computer and network consulting services provided on a recharge basis (system administration, desktop support, IT security consulting, firewall administration, etc.)

SOFTWARE LICENSING – Negotiates and manages software licenses and volume discounts for various agreements.

APPENDIX C: BIG DATA INFRASTRUCTURE EFFORTS ELSEWHERE

This section reviews sample models of Big Data infrastructure at other institutions. The purpose is to demonstrate that other universities are tackling these challenges and to provide context for how UC Davis might build its infrastructure.

1. UC San Diego

UCSD's Big Data model consists of three components.

Triton consists of two compute clusters available to researchers at UCSD and other institutions of higher education (on a fee basis). The first cluster contains 28 nodes combined with a large memory resource designed for data analysis on large data sets. The second cluster contains 256 nodes designed for parallel processing.

iDash (Integrating Data for Analysis, Anonymization and Sharing) is one of five National Centers for Biomedical Computing. The goal of the center is to develop new ways to gather, analyze, use and share vast, ever-increasing amounts of biomedical information. iDash provides computing services, algorithms, open-source software, and data storage and training to biomedical, clinical, and informatics communities at universities, medical schools, and hospitals nationwide.

Finally, the *RCI* (Research CyberInfrastructure) is a service which offers UC San Diego researchers the computing, network, and human infrastructure needed to create, manage, and share data in a manner that addresses federal sponsors' existing and new data management requirements. The program offers campus researchers facilities, storage, data curation, computing, colocation and networking services to facilitate their research using shared cyberinfrastructure services across campus. The UCSD Chancellor allocated an initial two-year start up budget of $3,000,000–$5,000,000 for five pilot projects to test the infrastructure and develop cost models. This first phase is under an RCI Oversight Committee, co-chaired by the San Diego Supercomputing Center director and the dean of the Medical School. They are now seeking $5,000,000 for 3 years of follow-on funding and to do assessment of the pilot phase.

2. University of Washington

The University of Washington has established the eScience Institute to take advantage of the unprecedented opportunities for discovery presented by the abundance of data that is now being produced by researchers and instruments in a variety of disciplines. The Institute was developed in a 'bottom-up model' driven by the needs of the research groups, and it is organized in such a way as to embed research scientists within different departments or programs, thereby enabling the transition of researchers in that field to data intensive applications. The eScience team provides computer science expertise, and individuals with backgrounds in a wide variety of academic disciplines help domain scientists apply the most appropriate technology to their research (e.g., with the application of computational methods, tools, and other resources for data-driven discovery, as well as proposal development). The Institute holds workshops and educational programs for graduate students and researchers and has had a broad influence across many areas of the university.

In addition to intellectual resources, the Institute provides physical infrastructure to support data-driven research, including a shared high-performance computer cluster, a file-based storage service that can be used to share data on and off campus, and long-term archiving of data that will be rarely accessed. Both the compute and storage services are connected to the campus and research networks at very high speeds. The eScience Institute has very close ties to the UW Information Technology team, who operates the storage and compute clusters.

3. Stanford University

In late 2012 Stanford's School of Humanities and Sciences and the School of Medicine announced a joint initiative to engage in interdisciplinary collaborations that will catalyze discovery in emerging fields of research, with a particular focus on the "information age of genomics." The Stanford Center for Computational, Evolutionary and Human Genomics will take advantage of the intersection between medicine, science, engineering, and the humanities to turn the translation of genomic data into scientific advances that can help promote health, agriculture and biotechnology. The center is open to all of the university's faculty and laboratories, and provides:

- Support for graduate and postdoctoral students;
- Support for small project grants, including student-initiated research;
- A computational genomics analysis service to support member labs and faculty, students and staff;
- Public outreach -- for example, during the first year, the center will present programs on "Genomics and Social Systems," "Agricultural, Ecological and Environmental Genomics," and "Medical Genomics"; and
- Consultation with academic institutions, industry, and government and nonprofit organizations to facilitate collaboration and transfer knowledge.

4. Purdue University

Purdue University has the Rosen Center for Advanced Computing (RCAC), which provides the campus with access to leading-edge computational and data storage systems, as well as expertise in a broad range of high-performance computing activities. RCAC is the research arm of the Information Technology at Purdue (ITaP), and it provides advanced computational resources and services to support Purdue faculty and staff researchers. RCAC also conducts its own research and development to enhance the capabilities of these resources. It provides services like DiaGrid and TeraGrid (for large-scale HPC), PURR (for research data management), HUBzero, a platform for building web sites that support scientific discovery, learning, and collaboration, and significant shared cluster computing infrastructure developed over several years through focused acquisitions using funds from grants, faculty startup packages, and institutional sources. These "community clusters" are now at the foundation of Purdue's research cyberinfrastructure. This Center has attracted programs like the Network for Computational Nanotechnology (NCN), which advances nanoscience and nanotechnology through online simulation and other resources on nanoHUB.org. nanoHUB has become a successful, scientific end-to-end cloud computing environment, hosting over 3,000 resources for research, collaboration, teaching, learning, and publishing.

5. Summary

The four models described above have certain features in common. The first is that they are very new, and are in the process of developing the services they will provide and the infrastructure needed to support those services. Indeed, UCSD is planning to assess its program to determine how best to provide those services to its clients. The second is the recognition that ongoing support of the infrastructure is critical to the functioning of Big Data services. Providing ongoing support enables the infrastructure to function, and indeed to evolve as the resources and needs of the users evolve. The third is that staff provides critical support for the infrastructure, assisting users to access and use the Big Data resources. At an academic institution, students will provide some of this support; as they graduate, new students will take their place. This means that the support staff training is ongoing, and must be supported to provide an infrastructure that non-computing experts (and, sometimes, experts) can use. Any infrastructure that UC Davis develops should have these characteristics.

APPENDIX D: CHARGE LETTER AND COMMITTEE MEMBERSHIP

UNIVERSITY OF CALIFORNIA, DAVIS

BERKELEY • DAVIS • IRVINE • LOS ANGELES • MERCED • RIVERSIDE • SAN DIEGO • SAN FRANCISCO SANTA BARBARA • SANTA CRUZ

OFFICE OF THE PROVOST AND EXECUTIVE VICE CHANCELLOR
ONE SHIELDS AVENUE
DAVIS, CA 95616
TEL: (530) 752-4964
FAX: (530) 752-2400
INTERNET: http://provost.ucdavis.edu

October 15, 2012
Revised October 16, 2012

Professor Nina Amenta, UC CITRIS Director, Computer Science, COE
Professor Matt Bishop, Computer Science, COE
Assistant Professor Graham Coop, EVE, CBS
Associate Vice Chancellor Paul Dodd, Interdisciplinary Research and Strategic Initiatives
Professor Jonathan Eisen, EVE, CBS
Professor Mike Hogarth, Department of Pathology, SOM
Associate Professor Ben Houlton, LAWR, CA&ES Professor
Ken Joy, Computer Science, COE
Professor Louise Kellogg, Geology, MPS
Professor Kenneth Kizer, Director of the Institute for Population Health Improvement, SOM Professor
Mike Kleeman, Civil and Environmental Engineering, COE
Professor Patrice Koehl, Computer Science, COE (Co-Chair)
Professor Ian Korf, Genome Center, MCB, CBS
Professor Kwan-liu Ma, Computer Science, COE
Professor and Director Richard Michelmore, Genome Center
Associate Professor Colin Milburn, English, HArCS
Professor Wolfgang Polonik, Statistics, MPS
Associate Professor Kim Shauman, Sociology, DSS
Vice Provost and CIO Peter Siegel, IET
University Librarian MacKenzie Smith, University Library
Professor Tony Tyson, Physics, MPS (Co-Chair)
Professor Bart Weimer, Population Health & Reproduction, SVM
Associate Professor Catherine Yang, GSM
Assistant Professor Huaijun Zhou, Animal Science, CA&ES

RE: Charge for the “Big Data” Implementation Committee

Dear Colleagues:

We as a campus and society are increasingly faced with unprecedented volumes of data in many different scientific and non-scientific fields. We therefore need to prepare our teaching and research programs as well as develop our infrastructure to accommodate this impending reality. There have been several whitepapers that address this issue and there are some on-going efforts on campus in this area; however, there is no overall coordination and little communication between these efforts and some major areas are not being addressed. In addition, there is an increasing awareness of the issue at the Federal level, including the President Obama’s “Big Data” Initiative[1]. We need to position ourselves quickly to take advantage of opportunities for funding through this and other impending initiatives. To address this need, I propose the establishment of the *“Big Data” Implementation*

[1] http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf

Committee to review the issue and make substantive recommendations for action. I write to request your service on this *ad hoc* committee.

The specific charges to this committee are:

- To review and update the recommendations from previous whitepapers, as well as materials from other campuses and agencies, relating to the handling and analysis of large datasets.

- To propose the path for developing strong, potentially disruptive academic research and teaching programs that address the opportunities and challenges presented by the Big Data revolution. In particular, the committee should identify both the fundamental elements of academic programs that are required as well as unique attributes that will distinguish efforts at UC Davis from those elsewhere.

- To consult with the campus at the levels of the Deans and departments as well as the faculty at large to coordinate activities related to Big Data so that the proposed plan addresses the diverse needs of the campus.

- To work with campus leadership to develop a comprehensive blueprint for the required intellectual resources (faculty and staff), along with the broad computational infrastructure requirements needed to support Big Data research and teaching needs at UC Davis. Recognizing the heterogeneity across the disciplines, we emphasize the need for advice on the key requirements and use cases for big data at UC Davis, not necessarily selecting particular technical architectures. For planning purposes, the committee might assume that an immediate follow-on implementation planning process, led by IET, will design or acquire the required services and architectures, based on the specific and broad requirements articulated.

- To identify imminent funding and recruitment opportunities that the campus should address. The need for this implementation plan is acute. Therefore, I ask that you complete your report to me by February 15, 2013.

Please accept my sincere thanks for your willingness to serve on this committee.

Sincerely yours,

Ralph J Hexter

Ralph J. Hexter
Provost and Executive Vice Chancellor

/mbm

c: Faculty Advisor Burtis
AEVC Mohr

APPENDIX E: ACKNOWLEDGMENTS

The Big Data Implementation Committee wishes to thank several individuals who made key contributions towards developing and drafting this report. We are grateful for the level of cooperation, assistance and expertise they provided.

- Andrew Connolly -- Professor, Department of Astronomy, University of Washington
- Rick Kiehl -- Professor and Chair, Department of Electrical & Computer Engineering
- Duncan Temple Lang -- Associate Professor, Department of Statistics
- Mark Redican -- Director, Communications Resources & Data Center Services, IET
- Babette Schmitt -- Chief of Staff, IET, & Support to the Big Data Implementation Committee
- Gabe Youtsey -- Interim Assistant CIO, IET