National Consortium for Data Science

Establishing a National Consortium for Data Science

Stanley C. Ahalt, Ph.D.
Renaissance Computing Institute
University of North Carolina at Chapel Hill

Contributing Authors

From the University of North Carolina at Chapel Hill:

Dan Bedard, *Renaissance Computing Institute*
Thomas M. Carsey, Ph.D., *The Odum Institute*
Jonathan Crabtree, *The Odum Institute*
Karen Green, *Renaissance Computing Institute*
Clark Jeffries, Ph.D., *Eshelman School of Pharmacy*
David Knowles, *Renaissance Computing Institute*
Hye-Chung Kum, Ph.D., *School of Social Work*
Howard Lander, *Renaissance Computing Institute*
Nassib Nassar, *Renaissance Computing Institute*
Arcot Rajasekar, Ph.D., *Data Intensive Cyber Environments Center*
Sidharth Thakur, Ph.D., *Renaissance Computing Institute*

Contents

Foreword

This white paper is not intended to be a static document, but rather a “request for comments.” As you read this paper, I encourage you to write in the margins, ask questions, and critique. Then, I would appreciate it if you would send me your notes so we can start a dialogue between RENCI and your organization. We believe the challenges of data science are too important and too difficult for a single organization to assume that it has all the answers.

I cannot emphasize this enough: **the challenges of data science are *critical* to the future of our academic institutions, our businesses, and our country.** So again, I encourage you to read the paper, send me your feedback, and let’s start working together to address the science of data.

Stan Ahalt
ahalt@cs.unc.edu

Executive Summary

The economic growth, national security, healthcare, food and energy production, and scientific progress of the Unites States depend upon a crucial, underdeveloped asset: **data**. As technological advances allow us to collect increasingly vast quantities of data, the analysis of massive data sets has emerged as a powerful new tool for scientific discovery and economic development. Those who harness the power of data will lead the 21st century.

However, efforts to fully capitalize on the potential of data are plagued by significant technical and societal challenges. These challenges constrain the application of data across all fields, and they cannot be addressed incrementally. To truly unleash the power of data and secure our nation's position as the global leader in the data-driven economy, the United States must invest in a sustained focus on the fundamental *science of data*.

The race is on. The European Union has already launched EUDAT, a major collaborative data management initiative. China continues to make major investments in advanced computing and is working on solutions to data science challenges. **Now is the time for the United States to establish a National Consortium for Data Science (NCDS)** to advance the application of data to solve grand-challenge problems, create jobs, protect national security, and improve our quality of life.

The proposed consortium will provide the intellectual and organizational infrastructure for top scientists and engineers to generate innovative solutions for data collection, storage, access, manipulation, and application. By cultivating the next generation of global data science leaders to advance data science and inspire radically new data management techniques for the benefit of science, government, and industry, the consortium will pay dividends for decades to come.

Data: The Currency of the Knowledge Economy

Every sector of our society—business, science, medicine, and government—is undergoing a historic transformation driven by the use of massive data collections, or "big data" (Maniyka, et al. 2011; Science 2011; Wood, et al. 2010). Today's technologies allow us to continuously generate vast streams of data about everything from customer transactions to genome sequences. These data collections hold tremendous potential to solve problems, create economic growth, and improve our quality of life (Horvitz & Mitchell 2010; Maniyka, et al. 2011).

According to *The Economist*, the data management and analytics sector is already worth more than $100 billion and is growing at almost 10 percent per year—about twice as fast as the software sector as a whole (Cukier 2010). Those who successfully manage and exploit data will drive innovation in commerce, scientific discovery, healthcare, finance, energy, and countless other fields. Those who harness the power of data will lead the 21st century. Data, says *The Economist*, "is the new raw material of business: an economic input almost on par with capital and labor."

Data is the currency of the knowledge economy. The economic success and security of the United States are inexorably bound to our ability to collect, manage, analyze, and understand data. Trillions of dollars of GDP are at stake (Orszag & Holdren 2010). **Yet, as we stand on the cusp of a data-driven economy, our ability to collect data has far outpaced our ability to manage and use it.** Researchers, governments, and industries cannot reap the full benefits of data because of limitations in our understanding of data and our technologies for handling it.

A special issue of the journal *Science* in February 2011, titled "Dealing with Data," highlighted critical data problems that hamper scientific research and applications in every field. One article on genomic data reported concerns about data access, security, and patient privacy and concluded that "a substantial leap in informatics capability and concomitant changes in the definition of data must take place to support movement of the field forward" (Kahn 2011). Another article on climate data reported that "a new paradigm of more open, user-friendly data access" is needed to work toward reducing society's vulnerability to climate

variability and change (Overpeck, et al. 2011). An article on sensor networks called for a "reinvention of sensor system design and signal processing theory" to cope with the overwhelming amount of data that sensors are collecting about physical and environmental conditions (Baraniuk 2011).

The problems of data are large, complex, and pervasive. This is not to say that innovative thinkers in industry and science have not made great strides in collecting and applying data. The strategic use of data has indeed opened doors to valuable business opportunities and new lines of scientific inquiry. However, many data solutions have been devised ad hoc to address niche needs, resulting in single-purpose systems that are difficult to adapt to other uses. In addition, fundamental data management problems remain, and a new class of problems is emerging as data sets grow ever larger.

A comprehensive, focused effort is needed in order to characterize, understand, and address fundamental data challenges. Other nations are already stepping forward to address the challenges of big data and take advantage of the opportunities it represents. In October 2011, the European Union launched EUDAT, a €16-million initiative to develop an international data management infrastructure over the next three years (EUDAT, 2011). Although China has not yet publicized a data science strategy, journal articles and Chinese investments in computing technology indicate a similar level of national interest in data.

To put data to work in novel ways and maintain a competitive edge in science and the global economy, the United States must develop and master *data science,* the systematic study of digital data. Only by applying established techniques of scientific investigation to digital data can the nation secure its place as the world leader in collecting, managing, analyzing, and using data, thereby positioning our businesses to take advantage of the power of data to drive economic growth. ***Therefore, there is a compelling case for a National Consortium for Data Science (NCDS)*** *to:*

- Engage a broad community of data science experts to identify key data science challenges,
- Coordinate data science research priorities, and
- Support the development of technical, ethical, and policy standards for data.

The NCDS will thus focus more attention on the necessity of fundamental data science research. Just as the NSF and DOE supercomputer centers programs allowed the nation's researchers and business leaders to exploit the power of high-performance computing for incredible scientific and economic gain, the NCDS will focus our nation on reaping the benefits of a data-rich society and foster the next generation of global data science leaders.

The NCDS is envisioned as a public-private partnership that will directly benefit universities, industry, and government by providing the intellectual and, in some cases, the physical resources needed to spur advances in data science. University researchers from many fields will be able to tap into the consortium's shared infrastructure to store and analyze large, complex sets of research data. These data sets will in turn serve as a test bed for top data scientists to develop fundamental data theories and principles and experiment with new methods of discovery, management models, data curation processes, and analytic methods.

The NCDS will establish close partnerships with leading private-sector firms to facilitate technology transfer and develop new products, and companies will have the opportunity to ensure their priorities are integrated into NCDS activities by funding projects or involving employees in the consortium to glean early insights into state-of-the-art data science and technology. Government agencies will be able to target their involvement in the NCDS to

Big Data Means Big Challenges

Big data is classified as data that exceeds the capacity of traditional systems. Data challenges can be "big" in terms of three characteristics, commonly known as the "Three V's":

Volume. Challenges arise from the amount of data that must be processed. For example, unfiltered, the data produced by the Large Hadron Collider would fill over 100,000 CDs per second (Beech 2010).

Velocity. Challenges arise from the need to process data within a certain timeframe. For example, financial trades must be analyzed in seconds to detect fraud or deliberately destabilizing activities.

Variety. Challenges arise from the heterogeneity of data needed to understand a situation. For example, an emergency manager may depend on data from weather sensors, traffic maps, and social media to conduct an orderly hurricane evacuation.

Despite its tremendous promise, we fail to fully capitalize on the potential of big data. Key challenges include:

- Storage and transmission capacity are not growing as quickly as our ability to produce data (for example, from DNA sequencing, Stein 2010).
- Available analysis tools often cannot efficiently process large data sets within practical time constraints.
- Effectively managing and using complex data sets requires expertise that spans information technology and domain-specific research areas.
- The cost of maintaining an infrastructure for big data threatens to shut out promising scientists and small businesses.
- Ethical standards, including privacy and security policies, must consider rapidly evolving technological capabilities.
- Without coordination, development efforts often produce piecemeal solutions.

investigate and solve critical scientific, economic, and security problems as they emerge. Across all these sectors, the NCDS will drive the field of data science forward by supporting the development of standards and best practices, curricular materials, and insights to support efficient, effective, and innovative uses of data.

NSF Supercomputer Centers: A Strategic Model

Twenty-five years ago, the NSF funded a modest set of investments in high-performance computing. At the time, there were about 40 supercomputers in the country, all but a few of them owned and operated by government agencies or major corporations. Many academic researchers found their computing needs rapidly outgrowing the computing capacity available to them, and few had access to supercomputers. Some researchers could only gain access to high-performance computing by traveling abroad.

In response, the NSF developed the Supercomputer Centers program, which established a shared supercomputing infrastructure that has given thousands of scientists easy access to supercomputing resources. That infrastructure has vastly improved researchers' ability to solve computationally intensive research problems and led to incredible scientific and economic advantages for the nation (Smarr 2009). In addition to providing the computing power researchers desperately needed, the centers have spurred innovation in computational software tools; provided training for students, researchers, and corporate partners in developing supercomputing applications; and forged private sector partnerships to apply the power of supercomputing to solve problems and advance business goals. The investment has proved to be one of the NSF's greatest successes.

Data science now presents a similar opportunity. If the United States can master the science of data and train a workforce that is facile in data science, it will reap long-term scientific, economic, and security benefits. Following the NSF Supercomputer Centers model, public funding should be applied to support national data science research for an extended period. In turn, this research will generate a long-term return on investment by propelling the nation to the forefront of data science and bringing data science advances to bear on research, business, and government goals (Hey, et al. 2009; Maniyka, et al. 2011; PCAST 2010; Wood, et al. 2010).

What Is Data Science?

Data science is the systematic study of digital data using scientific techniques of observation, theory development, systematic analysis, hypothesis testing, and rigorous validation. Data science entails an integrated view of the challenges of data and thus adopts a holistic approach to confronting these challenges.

The purpose of data science is to enhance the ability to collect, represent, measure, and apply data in order to describe, predict, and explain natural and social processes by:

1) Creating knowledge about the properties of large and dynamic data sets,
2) Developing methods to share, manage, and analyze digital data, and
3) Optimizing data processes for factors such as accuracy, latency, and cost.

The goals of data science include basic research and discovery, as well as applied research designed to inform decision-making for individuals, businesses, and governments.

Why a National Consortium? Facing the Big Data Impasse

Human innovation has brought us from an age of papyrus and clay tablets to punch cards and magnetic tape to spinning disks and solid state media capable of storing entire libraries on tiny chips. Along the way, we have devised increasingly sophisticated models and techniques for collecting, organizing, and analyzing data. In the 1940s, Clyde H. Coombs at the University of Michigan began using geometrical methods to analyze the structure of psychological data (Coombs 1964). In the 1960s, the development of direct-access storage devices began to allow us to store significant amounts of digital data, which in turn drove the development of hierarchical network database models. In 1970, Edgar Codd introduced the Relational Model, which underlies modern

Relational Database Management Systems (RDBMSs) such as Oracle and DB2. Recent efforts have produced NoSQL data stores, alternate data representations, and data lifecycle management strategies.

This rich history of data science and innovation has brought myriad economic and scientific gains. However, we have come to an impasse. Just as the enormous potential of big data is coming into clear focus, it has become apparent that existing solutions for data collection, storage, curation, and analysis cannot scale up to meet our growing needs.

Past advances in data science have worked well for handling the data sets of the past, which, for the most part, were relatively small, homogeneous, and did not typically require near-real-time analysis. Until recently, keeping pace with advances in data storage technology has generally required only modest tweaks to existing data management strategies. In general, new approaches to dealing with data have evolved in response to the specific needs of researchers, governments, and businesses as technologies changed, without an overarching framework to ensure solutions would remain robust when used for different disciplines or applications.

Existing data management solutions are not capable of handling the large, complex data sets that are now available, and the incremental improvements that have worked in the past will be incapable of achieving the sophisticated data management strategies and applications we require for the future. The large size of today's data sets is one important aspect of the problem (Loukides 2010), but it is not the only aspect. The challenges of big data stem from the expanding volume, velocity, and/or variety—the "Three V's"—of data. As the volume of data, velocity of data, and/or variety of data exceeds the capacity of our data infrastructure, it becomes impossible to ensure that a data set is complete and consistent; it cannot be stored, reused, or analyzed in entirety; and the full value of the data cannot be realized. Consequently, today's data sets present a level of uncertainty independent of the underlying reality the data represents. **Now, more than ever, we need an overarching framework for understanding and handling data.**

Our approach to data must shift from a paradigm of developing ad hoc solutions for idiosyncratic problems to a systematic study that generates an understanding of the production, management, and use of data. Understanding data requires the same scientific rigor as the systems described by the data. Data science should embody the four key elements of any scientific domain (King 1994). Thus:

1) The goal of data science is inference.
2) The procedures are public and subject to replication and peer review.
3) Conclusions are necessarily uncertain.
4) The content of data science is in the method(s).

The study of data must evolve into an interdisciplinary science that applies the scientific method to develop principles and theories that can guide new discoveries and innovations. **A sustained, targeted focus is needed to drive the field forward. Only through such focus can we achieve the data science discoveries and engineering techniques that will underpin the next great wave of global economic activity.**

Data Science: A Catalyst for Progress

To realize the full promise of data, data scientists must address data challenges in three dimensions: **data flow, data curation, and data analysis**. Each dimension presents unique problems with regard to both hardware and software, and each has attendant theoretical problems, including mathematical, statistical, engineering, and computer science challenges that require focused scientific investigation.

Major advances in data science require simultaneous consideration of all three dimensions in a unified, holistic approach. A National Consortium for Data Science will provide the framework necessary to advance data science, address these challenges, and spur innovations.

Data Flow

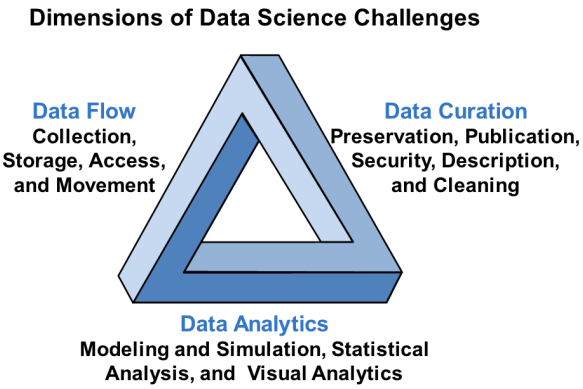
The first of the three dimensions of data science concerns factors intrinsic to the nature of data itself. This dimension includes issues relating to

the collection, storage, access, and movement of data—collectively termed "data flow." Every aspect of data flow is ripe for scientific exploration and further development as we struggle to manage and use big data.

Data Collection: Modern scientific instruments have made data collection a complex endeavor. The stereotypical image of a scientist in a lab coat recording measurements in a notebook has been replaced by images of sophisticated machinery rapidly harvesting vast quantities of data. For example, **in 2000 the Sloan Digital Sky Survey collected more data in its first few weeks of operation than had been collected in the entire history of astronomy.** The project has since collected more than 140 terabytes of data. The Large Synoptic Survey Telescope, due to come on line in 2016, will produce that amount of data every five days. Genomic analyzers now generate a petabyte (1,000 terabytes) of data per year. It is infeasible to store these vast data collections in their entirety; instead, system designers must use signal processing to extract data samples of manageable size**. Developing effective signal processing techniques for handling large data streams will be an active area of data science research.**

Data Storage: Data storage systems are typically composed of hierarchies of disk arrays and computer clusters which store volumes of data that exceed the capacity of individual devices. The two main technologies underpinning these systems are magnetic and solid state media. Each of these technologies has tradeoffs: solid state devices are more expensive, but are faster and consume less power; magnetic media can hold more data in a smaller amount of space. Often, these tradeoffs are only considered during deployment of a data system, yet data requirements are dynamic: a volume-intensive data set may at times require rapid analysis, becoming intermittently velocity-intensive. **Methodical experimentation and analysis are needed to characterize the tradeoffs between storage architectures in order to develop systems that can be reconfigured in real time.** The goal of this research will be data centers that are more scalable, more energy efficient, and less expensive to build and maintain.

Data Access: Once a physical storage medium is selected, designers must decide how to organize and store the data—a decision that



directly affects how the data are presented to and accessed by users.

The options for data organization and access are in a state of flux. In addition to the traditional flat file and Relational Database Management Systems (RDBMS), recent years have seen an explosion in "NoSQL" technologies. NoSQL products are intended to address performance and scalability problems in traditional databases by relaxing constraints such as data availability and consistency. However, NoSQL systems have yet to converge on a single storage strategy, storage format, or set of access mechanisms. A comprehensive set of guidelines is needed to help system architects match their workflow needs to the possible data access paradigms in order to inform their decisions about storage strategies. To that end, there is a need for data scientists to categorize workflow types, describe data access patterns and requirements, and document the performance and operational characteristics of various data storage technologies.

Data Mobility: As data sets grow ever larger, the ability to move them in a reasonable amount of time over commodity networks decreases. Currently, the industry standard is a 10 Gigabit-per-second link, which at its maximum rate takes 20 minutes to move a terabyte of data and two weeks to move a petabyte of data. Large-scale data users such as Google have expressed a desire for networks capable of moving data at 100 Gigabits-per-second, but 100 Gigabit links have not yet been widely adopted. Although network capacity is increasing, the amount of data to be moved is increasing at a faster rate. **Large data sets are at risk of being effectively isolated in place**. Data scientists can address this issue using several possible

The Promise of Data Science: Genomics

Opportunity. Biological and medical research are in the midst of a transformation made possible by the advent of low-cost high throughput genome sequencers capable of sequencing an entire genome in hours.

Impact. Applications range from studying human DNA for medical and anthropological purposes to studying microorganisms that affect our food supply to engineering synthetic organisms for agricultural and industrial production.

Challenges. While the promise of sequencing technology is great, the data challenges are enormous. Genomic laboratories can easily generate terabytes of data in a few months; the volume of data requires dedicated IT staff and hardware, as well as informatics experts to analyze and derive results from the data. Current data challenges include:

- Legacy databases, initially developed under stringent cost constraints, cannot scale to contain and manage larger data volumes.
- Compression algorithms that do not degrade data that may become useful later on need to be developed.
- The inability to easily share large data sets.
- Protecting private or sensitive data.
- Maintaining provenance of dynamic annotation metadata.
- Accelerating time-intensive data restructuring processes used to prepare data for analysis.

An Illumina Genome Analyzer produces ½ petabyte of data per year and costs a quarter of a million dollars. Such machines are becoming increasingly common as demand for genomics research continues to grow and sequencing costs plummet. *Image from Illumina, Inc.*

- Maintaining quality control as processes are automated and manual inspection diminishes.
- Building reference data sets that include all identified variations.
- Lack of tools that can manipulate, analyze, and visualize data at the terabyte scale.

The NCDS will engage genetics researchers and the broader IT and computer science communities to meet these data challenges and to accelerate genomic research.

strategies, including integrating computational functions into data centers, thereby reducing the need to move data; developing methods for breaking data sets into smaller, more mobile subsets; developing analytic techniques that can sample from large data sets; or finding ways to use existing networks more efficiently.

Data Curation

The second dimension of data science, data curation, concerns manipulating and recording data flow. Effective data curation enables analysis, the third dimension of data science, by adding context and making data reusable and discoverable (CIRSS 2012). Important areas of focus for future data science research efforts include data preservation, data publication, data security, data description, and data cleaning.

Data Preservation: Historical documents such as papyrus manuscripts and stone etchings are available today because the media used to record them required little active effort to maintain the fidelity of the data they contain. Preserving digital data is not so straightforward: hard disks have an estimated lifespan of three to six years, and even though DVDs and flash drives may last up to 100 years (ZDNet 2002), the technology to read them may not be available in the distant future. Technology emulation and information migration are the main techniques used to preserve digital data (Lee, et al. 2002), but these processes require significant sustained effort. Alternatively, rules

can be developed to determine which data should be archived based on the likely short and long term value of the data. Additional research is needed to reduce the economic burden of preserving large data sets.

Data preservation is further complicated by the fact that the analyses and processes that are applied to data often change it, resulting in data sets that are larger—or sometimes smaller—than the original collections. Processes that reduce data, for example by filtering a raw electrical signal before recording it, may inadvertently limit its utility for other disciplines. Preserving original, unprocessed data sets can leave the data set open for unanticipated secondary uses. For example, weather radar data has recently been used to track bird and insect migration patterns (Bruderer 1997). In other cases, processing results in data sets of vastly increased size. In such cases it may be preferable to store a record of the input data and a record of the processing system rather than the full results of the calculation. Preservation research is needed to develop optimal methods for retaining and archiving data.

Data Publication: For centuries, the scientific community has relied on journal articles and conference proceedings to share discoveries and give researchers credit for their work. Sharing scientific discoveries in this way facilitates peer review and accelerates research progress. Sharing digital data is useful for many of the same reasons. However, methods for publishing digital data are still under discussion (Callaghan, et al. 2011), and it is often necessary to protect data against misuse. New guidelines and methods for publishing and accessing data are needed to promote productive collaboration while protecting sensitive data.

Data Security: There are two facets to data security—data integrity and privacy protection. Digital data must be protected from corruption, loss, accidental errors, and intentional manipulation. This requires improvements in hardware and technology, but also in the development of policies for data curation. Methods already exist for auditing digital data sets to determine if they have been altered, but additional research and implementation techniques are required.

Data security also involves protecting data from misuse. For example, some smart grid data may reveal vulnerabilities in the power grid, and

The Promise of Data Science: Water Sustainability

Opportunity. A tremendous amount of water-related data exists in the form or sensor measurements, maps, and model outputs. Yet it is often difficult for water scientists to access and apply the data they need.

Impact. Understanding the dynamics of water is increasingly critical as population growth, climate change, and pollution place increasing burdens on clean water supplies. Enabling hydrologists, engineers, and urban planners to work with each other's data sets will play a pivotal role in addressing the grand challenge of water sustainability.

Challenges. Both technical and non-technical barriers impede progress in water science. A coordinated effort is needed to make water data and models interoperable. Key challenges include:

- Accessing diverse, heterogeneous data sets and metadata maintained by multiple local, state, and federal agencies; industry; environmental observatories; and individual researchers.
- Linking data sets and models from diverse fields to support the simulation of complex environmental systems and their interactions with humans.
- Transforming sensor and model data into information and knowledge products, documented with data provenance and error bounds to support additional analysis, modeling, and visualization.

Advancing data science is central to advancing water science, to understanding how climate change and population growth affect ecosystems and human well being, and to developing sustainable water management strategies. The NCDS will provide the infrastructure and the collaborative environment to help water scientists address their data challenges.

openly sharing that information could put vital infrastructure at risk (Boyer & McBride 2009). Consumer data, healthcare records, Facebook postings, and Google searches all document the actions and behaviors of people. Sensitive data can be protected through access control, encryption, or anonymization. However, even if individual data sets are anonymized, individuals often can be identified by combining data from multiple sources. One study suggests that 63% of Americans can be identified from their birth date, gender, and ZIP code (Golle 2006). Researchers have also developed statistical methods that can de-anonymize large digital data sets (Narayannan and Shmatikov 2008). We need to develop new techniques in both the technical *and* policy spheres in order to better protect privacy. As the White House report *National Strategy for Trusted Identities in Cyberspace* declares, "A secure cyberspace is critical to our prosperity… protecting it—while safeguarding privacy and civil liberties—is a national security priority and an economic necessity."

Data Description: Data is never self-explanatory. Every data set requires metadata to document the context of the data, typically including information about how the data has been processed, as well as when, why, and by whom it was processed. Metadata may also include information about the data set's formats and standards to help others read and use the data, replicate processing methods, control access, and retain and archive the data. Even with all of this information, however, there is no "Google" for navigating science data. Research is needed to refine how data should be described to best inform standards and formats, and to accelerate data mining processes.

In addition, data users in different scientific fields often use different vocabularies to describe the same concept or phenomenon. **Ontology mapping—a technique used to link disparate vocabularies used in different data sets—is critical to the ability to compare data across data sets and to enable researchers to use data from different fields.** Further data science research is needed to refine approaches to data description and ontology mapping.

Data Cleaning: Real-world data sets *always* contain some errors. Errors can be introduced at any processing stage, from the point of collection or data entry through the end points of analysis and reporting. Incorrect or inconsistent data can significantly distort the results of analyses (Hellerstein 2008). The saying, "garbage in, garbage out" holds true for data sets of all sizes and levels of complexity. Indeed, as data sets grow larger and more complex, the amount of "noise," as well as extra, missing, or incorrect information also increases. For example, many large data sets being created today actually combine data from several sources, which often use different collection methods, vocabularies, and database semantics. More research is needed to develop accurate and efficient processes for data cleaning—identifying and, where possible, correcting errors—to allow effective and responsible use of data.

Data Analytics

Data is generally not useful on its own. Using data to solve problems, answer questions, or make decisions requires some form of data analysis. Analytic processes transform data into forms such as statistical summaries, reports, and graphics that describe the data and catalyze new insights. Data analysis becomes more important as data sets grow larger and more complex. Data science research is needed to develop new analytical methods, such as techniques for modeling and simulation, advanced statistical analyses, and new approaches for visual analytics.

***Modeling and Simulation*:** Computer models use mathematical equations to emulate physical systems. By varying input parameters, researchers can use models to understand and predict how a system would respond to changes in the real world. For example, weather models use topological maps and previous meteorological measurements to predict future weather activity. Aircraft models simulate how new designs will perform before companies commit to costly construction processes. Creating models and using them to conduct simulations often requires large, complex data sets and sophisticated computations. Developing a new model requires test data for verification, and output data from simulations may be even more massive than the raw inputs. Important areas for future research include studying how data architecture affects the performance of models and developing new approaches to optimize the architecture of large data sets.

The Promise of Data Science: Building a Smart Electrical Grid

Opportunity. The U.S. power industry is deploying cutting-edge digital measurement and control technologies that will change the operation of the national electrical grid. These technologies, which include recording devices such as Phasor Measurement Units and Intelligent Electronic Devices in the transmission and distribution infrastructure and smart meters in homes and businesses, will result in an **unprecedented amount of data flowing through the system**.

Impact. Smart grid technologies will provide power system operators the information they need to diagnose and rectify power quality issues, which presently cost hundreds of billions of dollars annually in lost productivity. Smart grid devices will also permit a wider variety of generation resources (e.g., wind, solar, battery, micro-grids) to be connected to the grid with appropriate market pricing signals. Smart meters will help consumers make informed decisions about the energy they use.

Challenges. The smart grid will play a critical role in the future economic development of the United States. However, as the number of data sources has increased, new data challenges have emerged. The following challenges must be addressed in order to reap the full benefits of smart grid technology:

- Optimizing sensor placement, bandwidth utilization, and data stores for high velocity distributed data.
- Dynamically adapting network topology to minimize processing and transmission delay for high priority data.
- Sharing data to permit critical analysis while preventing misuse.
- Maintaining data provenance.
- Recording audit trails for post-mortem event reconstruction.
- Performing robust analysis on data sets that are incomplete and that contain errors.
- Presenting data in a format that helps system operators and consumers make decisions that increase the stability of the grid.

The NCDS will bring electrical engineers and data scientists together to study and solve smart grid data challenges to ensure the safety, reliability, and economic viability of the U.S. power system.

***Statistical Analysis*:** Statistical analysis helps researchers develop and test theoretical models based on the correlation of values within a data set. Common approaches include simulation-based methods, interdependent/matrix-valued data analysis, and unsupervised machine learning methods. However, these tools, which are critical for scientific progress, can become intractable on large data sets. Furthermore, conventional classification methods may categorically reject outliers, observations in experimental data that fall outside the distribution of a bell curve or other classical distribution. This exclusion of outliers as a statistical convenience often results in models that can successfully predict behavior within a test set but that fail when applied to novel real-world data.

The “nonparametric statistical methods” approach (Hollander & Wolfe1999) uses rankings of measurements, rather than the actual measurements, to classify data points. Nonparametric methods can be used to efficiently analyze data sets without discounting the significant role outliers may play in observable behavior. This approach has been used to efficiently and reliably determine causal relationships in genomic microarray experiments (Jeffries 2011). Additional research is required to further develop statistical methods that can be used to efficiently produce and test models for understanding complex systems and large data sets.

***Visual Analytics*:** A key challenge of dealing with large and heterogeneous data is the need to display data in a format that will help users understand it. Standard visualization approaches alone are often not sufficient to generate insight about complicated data. In such cases, visual analytics tools are extremely valuable for enabling users to make sense of the data and to make decisions based on the data.

Visual analytics (VA) is the science of using interactive visual interfaces to support analytical reasoning about data (Thomas & Cook 2006). VA is an interdisciplinary field that combines powerful analytical techniques with interactive visualization to detect and understand relationships between known quantities and stimulate new discoveries (Cook, et al. 2007). VA combines different technologies to enable exploration of massive data sets (Keim, et al. 2010), including data analytics, data and rule mining, pattern recognition, dimensionality reduction, and statistical analysis. VA also provides interactive visual interfaces to facilitate real-time exploration of large amounts of data. Finally, VA exploits human perceptual capabilities and the relative expertise of the user to support discovery of patterns in complicated data.

VA techniques are especially useful with complex data sets that may contain multidimensional data from a variety of domains that have different structural properties. For example, VA has been used to mine social media data from Facebook and Twitter to map keywords related to presidential election debates as a function of time, location, gender, and age (Hao, et al. 2011; Diakopoulos, et al. 2011; Diakopoulos and Shamma, 2010).

Spurring future advances in VA requires a multidisciplinary research approach that addresses both the technical challenges of retrieving and displaying data and the human challenges of perceiving, analyzing, understanding, and interacting with data.

The Impact of Data on Science

The use of big data has begun to fundamentally change how scientists conduct research and what they define as knowledge (Boyd & Crawford 2011). While the explosion of data offers substantial advantages, scientists must re-evaluate how basic principles of empirical analysis translate into the big data age. Foundational concepts like theory, sampling, representativeness, and accuracy still apply to the analysis of big data.

More data does not equal better data. If a sample is not representative, suffers from non-random gaps, or poorly measures a concept under study, then the analytic value of the data will be limited. Considering social networks, for example, not everyone Tweets, nor can we

The Promise of Data Science: Understanding Social Connections

Opportunity. Human activities increasingly leave digital traces, including data from government services, purchases, healthcare use, and social networks. Researchers who can make sense of these information sources stand to gain insight into many of the most challenging problems facing our society—healthcare, education, employment, welfare, economics, and the environment.

Impact. The potential of social data is enormous: One recent study estimates that the economic impact of integrated data on patient services is valued at $300 billion (Manyika, et al. 2011). Purchasing data allows merchants to target product advertisements to individual consumers. Social networking data is already being mined to identify national security threats.

Challenges. Personal data presents unique challenges. Trends evolve rapidly, and privacy and security are critical to maintaining public trust and the future utility of the data. Specific challenges of working with social data include:

- Managing the significant error and noise in human-generated data.
- Making data accessible and usable to researchers and policymakers, while protecting the privacy of the constituents.
- Integrating and accessing data maintained by multiple local, state, and federal agencies, as well as businesses.
- Building effective models that can transform data into information and knowledge, which are needed to support transparency and decision-making at local, state, and federal agencies.

The NCDS will facilitate the creation of a secure data infrastructure that can properly capture, curate, and maintain the flow of social data to advance research in the social, behavioral, and economic sciences.

assume that social networks as measured by Twitter data correspond to other types of social networks people might form. As Boyd and Crawford note, “a dataset may have many millions of pieces of data, but this does not mean it is random or representative” (Boyd & Crawford 2011). While some have claimed that the availability of massive data will supplant the need for theory—that the data will speak for itself—such a perspective may, in many circumstances, be misguided. Indeed, for some problems, drawing accurate and useful conclusions from big data may depend more on substantive and statistical theory, rather than less.

Data Ethics

In *Ethics of Big Data*, Kord Davis argues that the production of big data and the pressure to collect and analyze it push business operations deeper into people’s lives (Davis & Patterson 2012). This also applies to the operations of scientists and governments. This push poses significant ethical concerns that must be addressed. Davis categorizes the ethical issues of big data into four key areas:

Identity. How is the relationship between a person and his or her data established and verified?

Privacy. How is the relationship between a person and his or her data protected from disclosure?

Ownership. How much control is given to a person over his or her data? How much control is given to the collector or purchaser of the data?

Reputation. Can the data and the curator be trusted to protect the subject’s identity, privacy, and ownership?

The possibilities for ethical missteps with big data are profound. Although research organizations have established Institutional Review Boards to oversee the ethical conduct of human subject research, these boards are not well prepared to address the complex issues posed by big data analytics.

By providing a forum for public, private, and academic stakeholders to discuss, understand, and agree upon ethical standards, the NCDS will play a leading role in helping philosophers, ethicists, public advocates, and technologists establish a coherent code for the ethical use of large-scale digital data.

Integrating the Science of Data to Address Big Data Challenges

If the technical dimensions and philosophical issues of data science were completely independent of one another, existing scientific disciplines could solve all our data problems. Computer scientists, information scientists, mathematicians, and philosophers could study their domains and develop solutions in isolation. Indeed, scientists in these disciplines have made great strides in these areas, and their work has significantly improved our ability to

Table 1 Big data—in terms of volume, velocity, and variety—presents problems along all three dimensions of data science challenges.

	Flow	Curation	Analysis
Volume	Limited network bandwidth, limited storage space, need for specialized access methods	Cleaning, discovering relevant data points, long-term preservation	Computationally intensive calculations
Velocity	High network and storage latency, need for specialized access methods	Usefulness of data that ages quickly	Computational latency
Variety	Heterogeneous data and data formats require unique data stores	Heterogeneous policies for data management, discoverability of varied data types	Algorithms for processing heterogeneous data types

work with data. Yet the technical dimensions and philosophical issues are not conveniently separated along disciplinary lines. They are deeply interconnected, with significant areas of overlap and interaction among them. As a result, the approach used in one area may have immediate and long-term consequences in other areas. For example, if a data-driven system needs to perform an analysis only one time but the analysis must be done in three seconds or less, that requirement will influence the database structure, data retention policies, and analysis algorithms that are used.

The unique challenges posed by big data lie at the intersections of these dimensions, making an integrated, holistic approach to solving the challenges of data science increasingly vital (see Table 1). As data volume increases, new approaches are needed to store and transport large data sets, new methods are needed to clean and preserve the data, and new analytical tools are needed that can rapidly extract key information from the massive amount of data available. Meeting these needs will require integrated approaches that coordinate solutions across the three dimensions of data flow, curation, and analysis. Similarly, increases in data velocity and variety present challenges across all three dimensions and amplify the areas of overlap among them. An integrated discipline of data science will provide the overarching framework needed to develop systematic, robust data management solutions to meet these challenges**. The need for a coordinated focus on data science is not new; however, it has been made even more urgent by the unique problems presented by today's ever-expanding data sets.** The challenges posed by big data are intrinsic, multidimensional, and complex, and consequently must be addressed through a focus on data science that adopts a holistic, integrated approach.

The Vision for a National Consortium for Data Science

The incremental, ad hoc investigations that led to past achievements in data management will not be sufficient to usher us into the big data economy of the future. Big data presents unique challenges that require a new, multi-disciplinary approach to produce integrated theories, spur innovations, and address emerging problems in data flow, curation, and analysis. To maintain our position as the global leader in data science and allow U.S. researchers, businesses, and government to fully capitalize on the promise of big data, we need to establish a National Consortium for Data Science.

The central purpose of the consortium will be to bring the nation's top minds together to solve our most pressing data science challenges. Big data problems are large, pervasive, and constantly evolving—they will not be solved by individuals working in isolation. A consortium is needed to build a data science community that is greater than the sum of its parts by facilitating frequent, close interchange among data scientists and businesses, governments, and researchers in data-intensive domains. The work of the NCDS will be driven by real-world data problems and enable the development of interoperable, real-world data solutions.

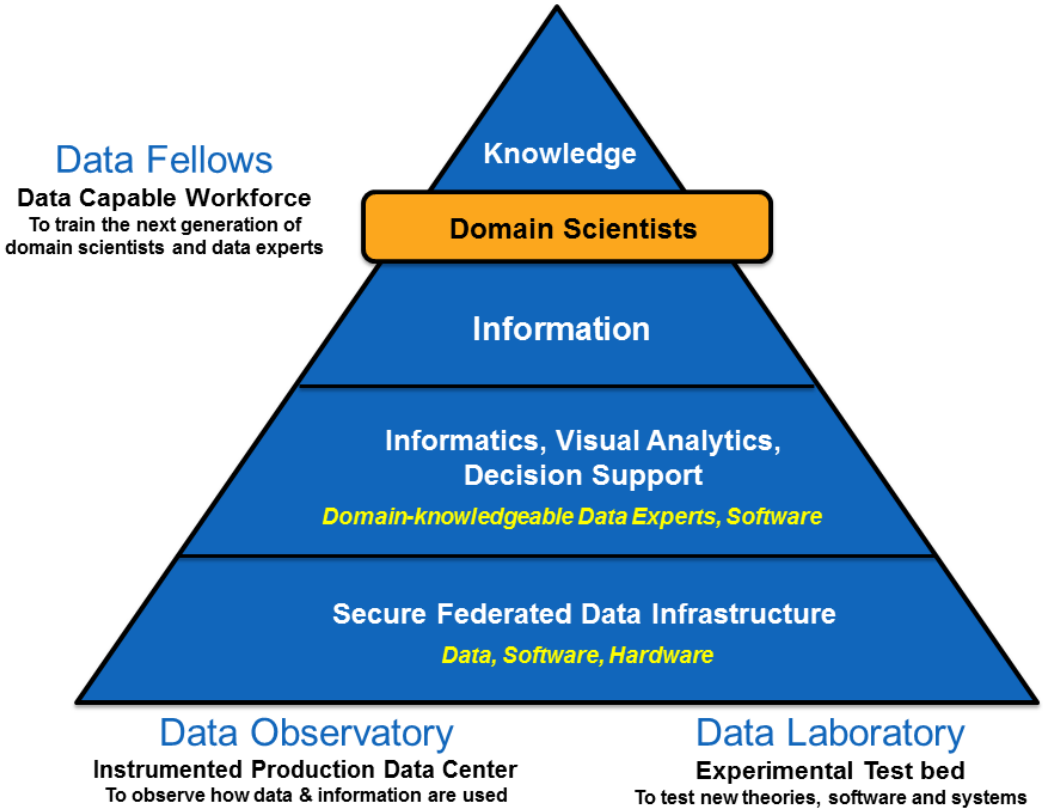
Three main components will be critical to the success of the NCDS. First, a ***Data Observatory*** will offer a shared and distributed infrastructure which provided members access to large sets of research data and gives researchers access to data to develop the theoretical underpinnings to inform data science advances. Second, a ***Data Laboratory*** will provide data science researchers access to emerging tools and the physical infrastructure they need to test radically new techniques for storing, sharing, analyzing, transforming, and visualizing data. Finally, a ***Data Fellows*** program will nurture a cadre of expert data scientists from universities, government, and industry to become tomorrow's global data science leaders.

To cultivate a culture of innovation and optimize its long-term impact, the NCDS will provide the organizational leadership needed to stimulate productive collaboration among the country's big data stakeholders, engage industry leaders to facilitate the commercialization and application of data science advances, and offer educational and workforce training and development opportunities to foster the growth of data science and a data-driven economy.

Data Observatory

The Data Observatory will give consortium members access to a robust distributed repository for data that will simultaneously

accomplish two goals. First, the observatory will provide a reliable data infrastructure that scientists in many fields can use to store, process, and access research data. For many researchers, the ability to tap into the consortium's shared infrastructure will allow them to work with larger, more complex data sets than can currently be handled at their home institutions. At the same time, NCDS data scientists will use the research data hosted by the Data Observatory to collect "data on data" by measuring how data is sensed, compressed, collected, stored, accessed, manipulated, modified, discarded, and transformed. Thus, the



Observatory will support the needs of the nation's data-intensive domain scientists while providing data scientists with an invaluable window into the properties of live data and the patterns of its use in scientific research. **No such observatory currently exists, and this unique facility will underpin the development of the field of data science.**

The Data Observatory will provide an environment for scientifically characterizing data, the information it contains, and the ways it is used. This environment will provide data scientists the deep understanding of data necessary for developing the theoretical branch of data science, which will guide experimental hypotheses and inform new data architectures. In addition, the critical mass of large, valuable data sets federated within the Data Observatory will encourage researchers from many scientific domains to coalesce around shared data models and interfaces for exchanging data. Finally, the highly interactive nature of the consortium will allow members to tap into the collective knowledge of the data science community to utilize the Data Observatory to its fullest potential. These factors **will make the Observatory a catalyst for transforming the data management field from a reactive paradigm based on commodity-driven hardware and re-purposed software to a proactive paradigm that will allow organizations and communities to manage collaborative data resources for maximum utility at minimum cost.**

To attract top researchers from institutions across the nation, the Data Observatory will give consortium members access to the collective infrastructures of the consortium members. A formal governance structure will provide a framework for selecting which research data

sets the Observatory will host, based on candidate projects' scientific merit and the goals of the NCDS. Careful monitoring will ensure the types of data accepted by the Observatory continue to support the goals and priorities of the NCDS. As the needs of the consortium members grow and evolve, members will discuss future infrastructure investments that would allow them to reach the next level in data science advances.

Data Laboratory

The Data Laboratory, a virtual laboratory, will provide a shared experimental resource for data scientists to imagine, design, test, and refine radically new techniques. Designed as a flexible, modular test bed for data experimentation, the Laboratory will allow data scientists to bring observations and theories developed in the Data Observatory to bear on new conceptual and physical models in all areas of data storage, movement, transformation, and visualization—including both hardware and software.

The motivating principal of the Data Laboratory will be to promote experimentation and innovation across a broad spectrum of data technologies and methods. Potential research questions for experimentation in the Data Laboratory include:

- How should data be optimally distributed between optical, magnetic, and electronic media?
- How should data be arranged at the bit level, and how should bits be grouped in order to form higher-order data? Are variable-length data structures viable alternatives to the "hexadecimal-heavy" formats currently used?
- How should data files be federated into larger groupings?
- How should data be transported in and out of storage media?
- How should data be arranged for delivery to newly developed interfaces, and how should data be optimally staged for access by operating systems and external input/output systems?
- How should heterogeneous data types be optimally arranged in variant file systems?
- How should data be secured at the bit level and how should sensitive content be protected?
- How should extremely large files or geographically distributed datasets be managed?
- How can real-time, complex data streams be utilized?
- How can we accurately detect and model data drift in dynamic data?
- How can error and uncertainty in data be managed?
- How can new tools for visualization and data-driven decision-making be implemented?
- Other questions not yet articulated.

In contrast to current, one-size-fits-all data storage architectures based on spinning disks and disk controller interfaces, the Laboratory will emphasize the use of open and highly modular hardware, software, and services so that scientists may combine and rearrange techniques at any point in a data processing workflow. This approach will give the Laboratory the inherent flexibility needed to drive successful new technologies, inventions, software, and data management methods.

Data Fellows

Facilitating collaboration among researchers across the country and around the globe will be the NCDS's primary organizing principle. Although most members will engage with the consortium largely from their home institutions, the Data Fellows program will offer members the opportunity to work across the consortium's many offices on summer, semester and year-long research sabbaticals. The program will allow scientists from academia, government, and industry to work with the consortium's virtual infrastructure directly and, at the same time, engage in targeted data science activities with other members of the NCDS. In addition, the Data Fellows program will cultivate an elite cadre of researchers to advance the consortium's goals as the NCDS becomes the pre-eminent global hub of data science.

Fellows will be selected based on project proposals in data-intensive domain science research, data science research, or data workforce development. They will enjoy access to the full suite of NCDS resources, including use of the Data Observatory to formulate theories and hypotheses and access to the Data Laboratory to develop and test new techniques. The NCDS infrastructure, combined with a

stimulating intellectual environment in which Fellows interact extensively with other top data science researchers, will give the Fellows critical insights into the rapidly-evolving field of data science, as well as the tools to apply those insights to drive innovations.

Upon returning to their home institutions, Fellows will disseminate the knowledge gained through their NCDS work to their colleagues and students. From the NCDS perspective, the continuous exchange, as well as the influx of new researchers and new ideas, will ensure NCDS research programs and educational activities stay relevant to the diverse communities the consortium will serve.

Other NCDS Activities

In addition to the activities of the Data Observatory, Data Laboratory, and Data Fellows, the NCDS will engage in several cross-cutting activities to ensure the consortium's research and development achieves maximum long-term impact. These activities include leading a range of collaborative activities to bridge disconnected research and stakeholder communities, establishing a task force to develop data standards and guidelines, partnering with private-sector firms to facilitate technology transfer, and developing educational and workforce development activities to foster future growth in the data science field.

Collaborative Activities: Transformative science requires meaningful collaboration across research and stakeholder communities. However, it can be difficult for researchers and industry professionals to keep tabs on research being done in other disciplines or at other organizations. The NCDS will serve its stakeholders by fostering an environment of frequent, productive collaboration. To create a unified data science community, the NCDS will:

- Organize an annual NCDS Conference on a problem that is critically important to the Data Science community. This annual "Data Science key challenge" problem will be expected to generate ongoing scholarship, and attract funding. The annual NCDS Conference will be the keystone event for the NCDS during its formative years.
- Organize additional workshops and research symposia for the consortium's researchers and other stakeholders. These events will provide opportunities for pivotal face-to-face discussion and stimulate new data science questions, ideas, and collaborations.
- Establish subgroups focused around particular projects or problem areas and hold frequent focus-area conference calls and teleconferences to plan research activities. Periodic interactive workshops will allow these subgroups to explore specific scientific problems in depth through short-term, hands-on activities.
- Develop websites and collaborative tools to assist researchers as they work together remotely. These tools will provide geographically-distributed researchers access to shared data sets, as well as the infrastructure needed to work effectively with collaborators at other institutions.
- Publish and distribute a newsletter to keep consortium members, industry partners, and government stakeholders abreast of NCDS projects and results, membership, and future activities.
- Establish regular meetings or other mechanisms to facilitate close communication between researchers and industry partners to ensure research activities continue to address real-world needs and support technology transfer.
- Engage international partners in NCDS collaborations and research. Close connections among U.S. and international stakeholders will allow consortium members to draw upon the knowledge and approaches being generated around the world and enhance the standing and visibility of U.S. data science at the global level.

Data Standards Task Force: As noted earlier, there is considerable confusion and under-development of standards and guidelines that govern how data is used, with respect to both technical and non-technical issues. In concert with its core research objectives, the NCDS will foster community-wide discussions aimed at developing and implementing data standards. Modeled on the Internet Engineering Task Force (IETF), the Data Standards Task Force will feature:

- Open participation in all standards committees and discussion forums with no formal membership or dues.

- A mission statement with a simple goal of producing policies to “make data work better.”
- Decisions reached by “rough consensus.”
- Policies published in formal documents.
- A hierarchical governance structure for managing working groups.
- A regular meeting schedule.

The task force will assist public, private, and academic researchers in developing best practices for technical aspects of data, such as storage and curation activities, as well as non-technical concerns, such as retention policies and data ethics.

The NCDS will assist the Data Standards Task Force by maintaining mailing lists, hosting national meetings, and devoting resources to support standards research and participate in task force proceedings.

Industry Partnerships and Technology Transfer: For advances in data science to truly benefit the nation, they must be put to use in practical applications. The NCDS will establish close partnerships with leading private sector firms to facilitate technology transfer and ensure the consortium’s overall strategic directions address the needs of industry. In addition, the NCDS will aggressively pursue commercialization of its intellectual property. A consortium model will enable industry partners to contribute intellectual capital to the consortium while maintaining appropriate protections for proprietary information and trade secrets. Intellectual property generated by consortium-supported research projects will be offered for license to consortium members with a right of first refusal before being promoted to the wider marketplace. Industrial partners will also be able to propose and fund targeted, proprietary NCDS research projects. Data science innovations that are not sponsored by the consortium or by individual firms will be marketed to potential licensees worldwide. When appropriate, the consortium will encourage the formation of spinout companies based on NCDS-developed technologies. All proceeds from the consortium’s technology transfer activities will be reinvested internally to continue to advance the mission of fostering new innovations in data science.

Workforce Development and Education: The NCDS will offer a comprehensive array of education and training programs to develop a workforce with the data skills needed to thrive in the knowledge economy. Because the NCDS will be developing the foundation of data science, it will be able to shape its educational programs in new and innovative ways. These programs will serve students and young professionals at the undergraduate, graduate, and postdoctoral levels; in addition, the NCDS will offer short courses and workshops for seasoned professionals in academia, government, and industry focusing on special topics in data science. Highlights of NCDS educational programming will include:

- Establishment of a graduate certificate program in data science. Designed as a complement to traditional graduate degrees in a range of scientific fields, the curriculum will focus on large-scale data management, computational analytic tools, and proper handling of sensitive data. Eventually, the curriculum will be expanded into full Masters and Ph.D. degree-granting curricula that will be available to U.S. institutions.
- Creation of a Data Science Summer Institute for doctoral students and postdoctoral researchers comprising weeklong or multi-week intensive courses and training.
- Establishment of a Diversity Graduate Fellows program. Through this program, doctoral students from traditionally underrepresented groups will be invited to conduct research at the NCDS headquarters for one year.
- Development of an undergraduate program in data science. With an end goal of offering an undergraduate minor in data science, the program will create and coordinate courses, provide mentoring for students, and stimulate student engagement with data science and scientists by organizing mini-conferences and other events.

Where feasible, these programs will use online modules and competency-based progress measures, as well as distance learning tools to allow people from around the country to take advantage of NCDS educational opportunities without having to leave their home institutions.

NCDS Structure and Governance

The NCDS will be governed by a Steering Committee that will determine the consortium’s long-range strategic goals and a team of directors that will oversee the consortium’s day-to-day operations. In addition, a National

Advisory Council will generate ideas for exploration and serve as a sounding board for potential new initiatives.

The Steering Committee will have 12-15 members representing data scientists, researchers in data-intensive domains, data science stakeholders, and industry partners. The committee will determine the high-priority research themes of the NCDS and assess the consortium's performance annually. In addition, the NCDS will advocate for ethical and responsible uses of data. The Steering Committee will ensure NCDS projects meet appropriate ethical standards and do not infringe on personal privacy or distort data for unethical purposes. Other functions will include fostering collaborations, promoting the NCDS among political and educational leaders, and facilitating technology transfer. The NCDS Director will serve as the administrative lead for the committee.

The NCDS leadership team will include the NCDS Director and Associate Directors responsible for managing the consortium's finances, personnel, facilities, and infrastructure. Other NCDS faculty and staff will include Data Fellows, as well as personnel responsible for administering the NCDS; maintaining and advancing infrastructure for the Data Observatory and Data Laboratory; developing key software tools, documentation, and outreach materials; and developing educational materials, including a formal data science curriculum and training program.

The National Advisory Council will serve in an advisory, rather than a decision-making, role. Its membership will include the NCDS Director, staff of members of Congress who serve on science and technology committees, and leaders from industry, private foundations, and universities. The Council will make recommendations related to the long-range vision for data science, help strengthen political support, facilitate technology transfer, and foster national and international collaborations.

Conclusion

Many of the scientific achievements and economic opportunities of the past few decades have been made possible by advances in the collection, manipulation, and application of data. We are now able to harvest vast quantities of data about everything from real-time brain activity to trends in housing prices; these data streams are opening doors to sweeping societal changes, powering everything from Twitter revolutions to early-warning systems for tsunamis and disease outbreaks.

Historically, advances in collecting and using data have been made incrementally, driven by the evolving needs of researchers, governments, and businesses. Clearly, this system has produced significant achievements. **However, we stand at the cusp of a new era.** Despite the enormous promise of big data, we struggle to exploit data to its full potential. Our ability to store, organize, manipulate, and apply data has not kept pace with our ability to collect it. As a result, researchers and businesses are drowning in a data deluge.

The United States needs a sustained, comprehensive effort to integrate the disparate research activities of data science. Establishing a National Consortium for Data Science will provide the necessary focus to position our scientists and business leaders to realize the full potential of big data.

The NCDS will serve the nation through the following goals:

- Advance data science by fostering collaboration among data scientists, software engineers, and domain scientists with data-intensive applications.
- Support data-intensive research and applications in many fields by providing the cyberinfrastructure and expertise to manage and analyze large data sets from multiple disciplines to meet the needs of researchers, businesses, and government.
- Promote the ethical and effective use of data by supporting community efforts to establish data standards and policies.
- Cultivate the next generation of data science leaders.
- Secure the nation's place as the global leader in data science and position U.S. businesses to apply the power of big data to drive economic growth.

Joining the National Consortium for Data Science will pay dividends for many decades to come. Unleashing the full power of big data will open up new lines of scientific inquiry and amplify our ability to observe, quantify, and understand our world. Data-enabled insights will in turn inform our approaches to broad societal

problems such as disease, poverty, and environmental degradation. Fully exploiting big data will also propel U.S. businesses to the forefront of the knowledge economy by allowing them to harness the power of data to make decisions and adapt to rapidly changing situations. Finally, investing in data science will help the U.S. maintain its position as a global leader in science and industry and protect and defend the safe, secure environment U.S. citizens and businesses need to thrive.

References

Baraniuk, R. G. (2011). More is less: Signal processing and the data deluge. *Science, 331*(6018), 717-717.

Beech, Martin. (2010) The Large Hadron Collider: Unraveling the Mysteries of the Universe. New York: Springer.

Boyd, D., & Crawford, K. (2011). Six Provocations for Big Data. *SSRN eLibrary*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1926431&

Boyer, W. F., & McBride, S. A. (2009). Study of Security Attributes of Smart Grid Systems–Current Cyber Security Issues: Idaho National Laboratory, USDOE, Under Contract DE-AC07-05ID14517.

Bruderer, B. (1997). The study of bird migration by radar part 2: Major achievements. *Naturwissenschaften, 84*(2), 45-54.

Callaghan, S., Lawrence, B., Pepler, S., Jones, C., & Matthews, B. (2011). Citation and Peer Review of Data: Moving Towards Formal Data Publication. *International Journal of Digital Curation, 6*(2).

CIRSS (Center for Informatics Research in Science and Scholarship). (2012). Data Curation Education Program Retrieved February 12 2012, from http://cirss.lis.illinois.edu/CollMeta/dcep.html

Cook, K., Earnshaw, R., & Stasko, J. (2007). Guest Editors' Introduction: Discovering the Unexpected. *Computer Graphics and Applications, IEEE*, *27*(5), 15 –19. doi:10.1109/MCG.2007.126

Coombs, C. H. (1964). *A theory of data*. Oxford, England: Wiley.

Cukier, K. (2010, February 25). Data, data everywhere. *The Economist*.

Davis, K. & Patterson, D. (2012). Ethics of Big Data: Balancing Risk and Innovation.O'Reilly Media.

Diakopoulos, N., & Shamma, D. A. (2010). Characterizing Debate Performance via Aggregated Twitter Sentiment. *Conference on Human Factors in Computing Systems (CHI)*.

Diakopoulos, N., Naaman, M., Yazdani, T., & Kivran-Swaine, F. (2011). Social Media Visual Analytics for Events. *Social Media Modeling and Computing* (pp. 189–209).

EUDAT. (2011). "EUDAT: European Data Infrastructure." Retrieved February 28 2012, from http://www.eudat.eu/fact-sheet

Golle, P. (2006). Revisiting the uniqueness of simple demographics in the US population. *Proceedings of the 5th ACM workshop on Privacy in electronic society*, WPES '06 (pp. 77–80). New York, NY, USA: ACM. doi:10.1145/1179601.1179615

Hellerstein, J. M. (2008). Quantitative Data Cleaning for Large Databases: UC Berkeley.

Hao, M., Rohrdantz, C., Janetzko, H., Dayal, U., Keim, D. A., Haug, L., Hsu, M.-C., et al. (2011). Social Media Visual Analytics for Events. In S. C. H. Hoi, J. Luo, S. Boll, D. Xu, R. Jin, & I. King (Eds.), *Social Media Modeling and Computing* (pp. 189–209). Springer London. doi:10.1109/VAST.2011.6102472

Hey, A. J. G., Tansley, S., & Tolle, K. M. (2009). *The fourth paradigm : data-intensive scientific discovery*. Redmond, Wash.: Microsoft Research.

Hollander, M., & Wolfe, D. A. (1999). *Nonparametric statistical methods*. Wiley.

Horvitz, E., & Mitchell, T. (2010). From Data to Knowledge to Action: A Global Enabler for the 21st Century *A Computing Community Consortium White Paper*: Computing Community Consortium.

Jeffries, C. D., Fried, H. M., & Perkins, D. O. (2011). Nuclear and cytoplasmic localization of neural stem cell microRNAs. *RNA*, *17*(4), 675–686. doi:10.1261/rna.2006511

Kahn, S. D. (2011). On the future of genomic data. *Science, 331*(6018), 728-728.

Keim, D. A., Kohlhammer, J., Ellis, G., & Mansmann, F. (2010). *Mastering The Information Age-Solving Problems with Visual Analytics*. Florian Mansmann.

King, G., Keohane, R. O., & Verba, S. (1994). *Designing social inquiry: scientific inference in qualitative research*. Princeton University Press.

Lee, K. H., Slattery, O., Lu, R., Tang, X., & McCrary, V. (2002). The state of the art and practice in digital preservation. *Journal of Research of the National Institute of Standards and Technology, 107*(1), 93-106.

Lee, K., Brownstein, J. S., Mills, R. G., & Kohane, I. S. (2010). Does Collocation Inform the Impact of Collaboration? *PLoS ONE*, *5*(12), e14279. doi:10.1371/journal.pone.0014279

Loukides, M. (2010). What is Data Science? *O'Reilly Radar*: O'Reilly Media, Inc.

Maniyka, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity: McKinsey Global Institute.

Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. *Security and Privacy, 2008. SP 2008. IEEE Symposium on* (pp. 111–125).

Orszag, P. R., & Holdren, J. P. (2010). Appendix A: Office of Management and Budget and Office of Science Technology Policy *Science and Technology Priorities for the FY 2012 Budget, Memorandum for the Heads of Executive Departments and Agencies*.

Overpeck, J. T., Meehl, G. A., Bony, S., & Easterling, D. R. (2011). Climate data challenges in the 21st century. *science, 331*(6018), 700-700.

PCAST (President's Council of Advisors on Science and Technology). (2004). Revolutionizing health care through information technology *Report to the President and Congress*.

PCAST (President's Council of Advisors on Science and Technology). (2010). Designing a Digital Future: Federally Funded Research and Development in Networking and Information Technology *Report to the President and Congress*.

Science. (2011). *Dealing with Data* (Vol. 331).

Smarr, L. (2009). *The Good, the Bad and the Ugly: Reflections on the NSF Supercomputer Center Program*. Position paper submitted to NSF's Future of High Performance Computing Workshop, December 2009. Retrieved from http://www.calit2.net/newsroom/rss.php?id=1632

Stein, L. D. (2010). The case for cloud computing in genome informatics. [10.1186/gb-2010-11-5-207]. *Genome Biology, 11*(5) 207-207.

Thomas, J. J., & Cook, K. A. (2006). A Visual Analytics Agenda. IEEE Comput. Graph. Appl., 26(1), 10–13. doi:http://dx.doi.org/10.1109/MCG.2006.5

Wood, J., Andersson, T., Bachem, A., Best, C., Genova, F., Lopez, D. R., . . . Hudson, R. (2010). Riding the wave: How Europe can gain from the rising tide of scientific data.

ZDNet. (2002). Tech Guide: Storage media lifespans. Retrieved from http://www.zdnet.com.au/tech-guide-storage-media-lifespans-120269043.htm