REALIZING THE POTENTIAL OF DATA SCIENCE

Final Report from the National Science Foundation Computer and Information Science and Engineering Advisory Committee Data Science Working Group

Francine Berman and Rob Rutenbar, co-Chairs
Henrik Christensen, Susan Davidson, Deborah Estrin, Michael Franklin, Brent Hailpern, Margaret Martonosi, Padma Raghavan, Victoria Stodden, Alex Szalay

December 2016

I. INTRODUCTION & SUMMARY OF RECOMMENDATIONS

Much has been made of the rise of *digital data* as a driver that is advancing virtually every intellectual endeavor. In the research community, *data-driven* discovery is extending fundamental approaches that started with *observational* discovery and *theoretical* discovery, and has embraced the quantitative assets of the Information Age with *computational* discovery. In commerce, digital data often serves as a disruptive technology – fundamentally changing our ability to create value from information and to understand, interpret, and respond to the needs of customers and clients. The rise of a new paradigm – *Data* to *Knowledge* to *Action* – implies the use of data for decision making by *both* humans and machines, and creates novel opportunities as well as challenges.

It is not too extreme to say that *data is changing everything*. As a result, there is an emergence of a new field – *Data Science* – that focuses on the processes and systems that enable the extraction of knowledge or insights from data in various forms, either structured or unstructured. In practice, Data Science has evolved as an interdisciplinary field that integrates approaches from data analysis fields such as Statistics, Data Mining, and Predictive Analytics drawing on diverse observational domains. Of particular interest for this report is the deep connection between Data Science and Computer Science; as noted recently in *Forbes,* "[Data Science is] the story of the coupling of the mature discipline of statistics with a very young one–computer science."[1]

The ability to manipulate data and understand Data Science is becoming increasingly critical to current and future discovery and innovation. McKinsey predicts that data-driven technologies will bring an additional $300 billion of value to the U.S. health care sector alone, and by 2020, 1.5 million more "data-savvy managers" will be needed to capitalize on the potential of data, "big" and otherwise.[2] The rise of data is creating similar opportunities/challenges in fundamental science. For example, a 2015 study in *PLOS Biology*[3] asked the question, "Is Big Data Astronomical or Genomical?," comparing predictions to 2025 for the scale of genomic data, observational astronomy data, and commercial YouTube and Twitter data. The study concludes that genomics data "is either on par with or the most demanding of [these] domains … in terms of data acquisition, storage, distribution, and analysis." A recent paper in *Science*[4] observes that the social sciences are being similarly transformed, "A field is emerging that leverages the capacity to collect and analyze data at a scale that may reveal patterns of individual and group behaviors."

In light of the importance of data for research and commerce, the development of Data Science is increasingly crucial for the research and education community. How do we train a workforce of professionals who can use data to its best advantage? What should we teach them? How do we advance the field of Data Science so it can support the increasing role of data in all spheres? What can the National Science Foundation (NSF) do to help maximize the potential of Data Science to drive discovery and decision making, and address current and future needs for a workforce with Data Science expertise?[5]

---

[1] Press, G. (May 28, 2013). A Very Short History of Data Science. *Forbes*. Retrieved from http://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#5b07c37469fd

[2] Manyika, J. et al. (2011). *Big data: The next frontier for innovation, competition, and productivity.* McKinsey Global Institute. Retrieved from http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation

[3] Stephens, Z. et al. (2015). Big Data: Astronomical or Genomical?. *PLOS Biology, 13*(7): e1002195. doi: 10.1371/journal.pbio.1002195

[4] Lazer, D. et al. (2009). Life in the network: the coming age of computational social science. *Science,* 323(5915): 721-723. doi: 10.1126/science.1167742

[5] For example, NSF's program in Transdisciplinary Research in the Principles of Data Science (TRIPODS), takes a step towards identifying and establishing fundamental principles that need to be imparted to all data scientists. See https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505347.

It is this last question that motivated the Advisory Committee of the Computer and Information Science and Engineering (CISE) directorate of NSF to charter a Working Group in 2015 to examine questions surrounding the emergence of Data Science. The high level charge to the Working Group was:

> *Imagining the future for Data Science, articulate a vision for where the field might be 10-15 years hence, and identify the steps that NSF CISE specifically (but others as well) could take now and in the future – consistent with NSF's mission in the areas of research, education, workforce development, economic competitiveness, national security and more – to help realize this vision.*

This Working Group and its charge are timely. The rise of data is having enormous impact on the full range of science and engineering endeavors that comprise NSF's core mission. Data Science will advance both discovery and learning efforts within the NSF research and education community and impact the kind of infrastructure needed to support current and future efforts. It will also create new challenges. How do we make best use of datasets used for research and education? What kind of infrastructure will be needed to make data available, accessible, and useful? How will we ensure that data-derived findings are unbiased, reliable, and verifiable? How will NSF and its community address the challenges of data stewardship and sustainability?

If NSF can help foster the evolution and development of both Data Science and Data Scientists over the next decade, we can begin to meet the potential of Data Science to drive new discovery and innovation, and transform the Information Age into the Knowledge Age. This should include not only a focus on fundamental Data Science, but also *translational* efforts to move ideas from research to practice across the broadest landscape of commercial applications.

This short report is organized as follows: Section II focuses on the life cycle of data – from birth to death or immortality, especially as it is used for scholarly research – as a framing mechanism for understanding the diverse components of Data Science. Section III describes opportunities to evolve Data Science in three critical areas: research (Section III.A), education, and training (Section III.B), and supporting infrastructure (Section III.C) to meet its potential over the next decade. Section IV considers future areas of growth and innovation that will evolve Data Science, and whose evolution can be driven now by a strategic NSF agenda.

The following is a summary of the Strategic Recommendations for NSF in this report. Note that these Recommendations build on an existing and important collection of programs within CISE and throughout NSF that already support various areas of Data Science and that should be continued. It has been richly rewarding to develop these from a CISE perspective and our Recommendations are couched with that specific purview. It will be invaluable to complement these with insights from other NSF directorates and offices. The Recommendations herein are meant to highlight the importance of key areas in Data Science but should not be considered the only areas for investigation nor the entirety of a given agenda.

Strategic Recommendations for NSF

Recommendations for a National Data Science Research Agenda

R.1. CREATE DATA SCIENCE RESEARCH CENTERS. Focus/re-focus a set of subsequent calls for Science & Technology Centers (STCs) and Engineering Research Centers (ERCs) as *Data Research Centers*.

> *Utilize key Data Science challenges as the thematic focus of the Centers. Encourage proposers to take a comprehensive approach that uses the full data life cycle as a central organizing principal and value proposition in addressing these challenges.*

R.2. INVEST IN RESEARCH INTO DATA SCIENCE INFRASTRUCTURE THAT FURTHERS EFFECTIVE DATA SHARING, DATA USE, AND LIFE CYCLE MANAGEMENT. Develop programs that focus attention on critical problems (privacy, inference, provenance, etc.) that remain obstacles to the use of data at scale.

> *Research outcomes should ultimately be translatable to infrastructure that enable access to data in ways that: (i) preserve privacy and other commitments made when collecting the data; (ii) enable researchers to make "unbiased" inferences, or understand potential biases in the data and other data use challenges; (iii) support reproducibility; (iv) support access, provenance, sustainability, and other life cycle challenges; and (v) support research into new hardware/software infrastructures needed to support Data Science research.*

R.3. SUPPORT RESEARCH INTO EFFECTIVE REPRODUCIBILITY. Develop research programs that support computational reproducibility and computationally-enabled discovery, as well as cyberinfrastructure that supports reproducibility.

> *Potential research efforts may focus on, for example, mechanisms to extend validation, verification and uncertainty quantification to include reproducibility; software standards (creation, test, curation, etc.); tools for sharing and verifying queries on confidential data; tools to understand links between decision-oriented models and their training data (e.g., emerging artificial intelligence models); etc.*

R.4. FUND RESEARCH INTO MODELS THAT UNDERLIE EVIDENCE-BASED DATA POLICY AND DECISION MAKING. Invest in the development of models that can be used to support data-related policy, regulation, and strategic investment.

> *For example, investments in open access efforts should be supported by models that link data value, data cyberinfrastructure, and data investments. This would provide evidence for transparent policies and processes for exploring, for example, which digital artifacts might require additional resources to be invested in preservation, and how to make such decisions at the levels of the agency, enterprise, and institution.*

R.5. EXPAND FUNDING INTO DEEP LEARNING, SMART ENVIRONMENTS, AND OTHER ARTIFICIAL INTELLIGENCE-EMPOWERED AREAS AND THEIR USE IN DATA-DRIVEN APPLICATIONS. Continue to invest in both the foundations of "embodied intelligence" research and in their use by applications and within domains.

> *Encourage research that links to new domain applications, new cyberinfrastructure, and new curricula for Data Science.*

Recommendations for a National Data Science Education & Training Agenda

E.1. SUPPORT THE DESIGN AND DEVELOPMENT OF DATA SCIENCE PEDAGOGY AND CURRICULA. Make data a central focus for large cross-disciplinary education research centers, courses, complete curricula, and modes of pedagogy.

> *Include curriculum foci on Data Science research areas, social issues around data stewardship and use, and data life cycle components. Create opportunities across NSF directorates and offices to collaborate on Data Science education [e.g., Mathematical and Physical Sciences; Social, Behavior, and Economic Sciences (SBE), Biological Sciences, and Geosciences]. As two examples, NSF's Critical Techniques, Technologies and Methodologies for Advancing Foundations and Applications of Big Data Sciences and Engineering (BIGDATA) program might be naturally evolved or extended to include such activities, and NSF's Big Data Regional Innovation Hubs (BD Hubs) program could be additionally resourced to focus on effective workforce education and training as part of its core mission to develop big data sources, techniques, and shared infrastructure.*
>
> *In addition, support not only Science, Technology, Engineering, and Mathematics (STEM), but also non-STEM education opportunities in Data Science. For example, consider support for educational research on infusing statistical and machine learning "thinking" into business, government, and management programs, so that the results of big data can be effectively used by business leaders to make decisions.*

E.2. TARGET EXISTING OR NEW PROGRAMS TO DEVELOP DATA SCIENCE CURRICULA AT EPSCOR AND MINORITY-SERVING INSTITUTIONS. Include curricula and training programs that assist students at participating institutions to be competitive for Data Science internships in the private sector and fellowships in the academic sector.

> *Focus and support for institutions at all tiers is critical to develop the best and the brightest Data Science workforce and to make Data Science an inclusive area for all.*

E.3. ENCOURAGE "DATA INCUBATOR PROGRAMS" OR OTHER PRIVATE/PUBLIC PARTNERSHIPS THAT PROVIDE STUDENTS/FACULTY ACCESS AND OPPORTUNITIES FOR ENGAGEMENT WITH FACULTY/INDUSTRY/NON-PROFITS WITH REAL PROBLEMS. Develop programs that train a sophisticated Data Science workforce that is more prepared to address Data Science challenges at the frontier of innovation.

> *NSF's Industry/University Cooperative Research Centers Program (I/UCRC) is a start in this direction, and could be expanded, e.g., via the NSF Big Data Regional Innovation Hubs: Establishing Spokes to Advance Big Data Applications (BD Spokes) program.*

E.4. SUPPORT PhD AND POSTDOC FELLOWSHIPS IN DATA SCIENCE. These are important to address current and future research and workforce needs.

> *Fellowship programs could be supported through NSF Data Science programs or as an extra year in other programs with a plan for training in Data Science. For example, Data Science was an emphasis area for the first round of the NSF Research Traineeship program.*

Recommendations for National Data Infrastructure that Supports Data Science

I.1. SUPPORT THE CREATION, ACQUISITION, AND PUBLIC DEPLOYMENT OF A BROAD PORTFOLIO OF REALISTIC, STATE-OF-THE-ART, AT-SCALE DATASETS FOR ACADEMIC RESEARCHERS. Create collections of educational datasets and materials and make them available via publicly accessible repositories, libraries, and stewardship environments. Link them to the publications they support.

> *Creation of this fundamental infrastructure will drive increased inter-sector relationships (particularly with the private sector) to acquire and deploy datasets useful for scholarly research and Data Science educational programs.*

I.2. DEVELOP AND DEPLOY DATA SCIENCE INFRASTRUCTURE NEEDED TO SUPPORT CUTTING-EDGE RESEARCH AND EDUCATION. Invest in both national and institutional infrastructure to support emerging Data Science research and education programs.

> *Focus on low-barrier-to-access, representative, and sustainable Data Science infrastructure that supports research efforts and coursework. Initiate strategic and committed public-private partnerships that can help build representative at-scale infrastructure in the academic community for Data Science research and education.*

I.3. CREATE A NATIONAL EFFORT TO WORK WITH INSTITUTIONAL LIBRARIES AND DOMAIN REPOSITORIES TO DEVELOP SUSTAINBLE MODELS OF DATA STEWARDSHIP AND PRESERVATION OF VALUED SPONSORED RESEARCH DATA. Work with the community to develop and ramp up the use of sustainable institutional economic models that enable participating organizations to provide open access research data stewardship options to the community over the long term.

> *Evaluate outcomes thus far from NSF programs like Sustainable Digital Data Preservation and Access Network Partners (DataNet) and Data Infrastructure Building Blocks (DIBBS) and use lessons learned to inform an effective program.*

I.4. ESTABLISH BEST PRACTICE GUIDELINES FOR DETERMINING THE BALANCE OF DATA SCIENCE INVESTMENT THAT SHOULD BE FOCUSED ON RESEARCH VS. ENABLING INFRASTRUCTURE. Look to other organizations, institutions, and sectors to establish a rough guideline about the distribution of investment to maximize the effectiveness of research and education efforts.

> *To capitalize on infrastructure investments, encourage proposers to utilize existing useful data infrastructure when possible and to clearly articulate the value added for new infrastructure efforts in proposals.*

I.5. SUPPORT EFFORTS TO GATHER KEY INFORMATION ABOUT THE SCALE AND STATE OF SPONSORED RESEARCH DATA ACCESS AND STEWARDSHIP TO INFORM EVIDENCE-DRIVEN INVESTMENT AND POLICY. Gather statistics for sponsored research that describe the amount, characteristics, and stewardship of the data and allow linking between data and scholarly publication, perhaps through vehicles such as Fastlane or Research.gov.

> *Relevant characteristics may include amount and type of data generated, where the data is hosted, whether the data is available to a broader community, what publications the data is associated with, etc. Make this information available through the National Science Board Science and Engineering Indicators or a similar publication.*

Recommendations for New Data-Driven Scenarios

N.1. STRENGTHEN INTERNET OF THINGS (IoT)-RELATED DATA ECOSYSTEM. Focus on data aspects of IoT in critical areas such as data security, data privacy, and smart systems, as well as infrastructure supporting governance, policy, ethics, etc.

> *Broaden efforts to include consideration of ethics and data infrastructure needed for social governance of data and data policy. Provide leadership for defining and studying IoT data governance and social issues (perhaps in collaboration with SBE or through initiation of National Academies of Sciences, Engineering, and Medicine studies); develop pilot environments that can prototype prospective policies and technologies. Leverage programs and efforts like Smart and Connected Communities, Secure and Trustworthy Cyberspace, and others, to ensure that IoT is a driver for new research discovery as well as commercial innovation.*

N.2. SUPPORT FUNDAMENTAL RESEARCH ON RADICAL HARDWARE AND SOFTWARE SYSTEM ARCHITECTURES THAT TARGET EMERGING DATA-INTENSIVE TASKS. Focus on research that promotes innovative data analysis methods for next-generation scenarios and the evolution or re-thinking of the fundamental computing platforms on which they might best be executed.

> *The impending end of Moore's Law[6] and the rise of big data are already pushing data-oriented tasks onto GPUs, field-programmable gate arrays (FPGAs), accelerators, etc. Support fundamental research on radical hardware and software system architectures that target emerging data-intensive tasks. Leverage relevant efforts from the National Strategic Computing Initiative (NSCI).[7]*

[6] Conte, T. (August 8, 2016). Whistling Past the Graveyard: What the End of Moore's Law Means to All of Computing [Web log post]. Retrieved from http://www.cccblog.org/2016/08/08/whistling-past-the-graveyard-what-the-end-of-moores-law-means-to-all-of-computing/

[7] Exec. Order No. 13702, 3 C.F.R. (2015). Retrieved from https://www.whitehouse.gov/the-press-office/2015/07/29/executive-order-creating-national-strategic-computing-initiative

II. FROM CREATION TO INNOVATION – THE DATA LIFE CYCLE

An essential fact about data is that they never exist in a vacuum. Like biological organisms, data has a *life cycle*, from birth, through an active life, to "immortality" or some form of expiration. And like a living and intelligent organism, they survive in an *environment* that provides both physical support and existential meaning. Critically, software is always associated with data, for hosting, access, organization, analysis, and all parts of the data life cycle.

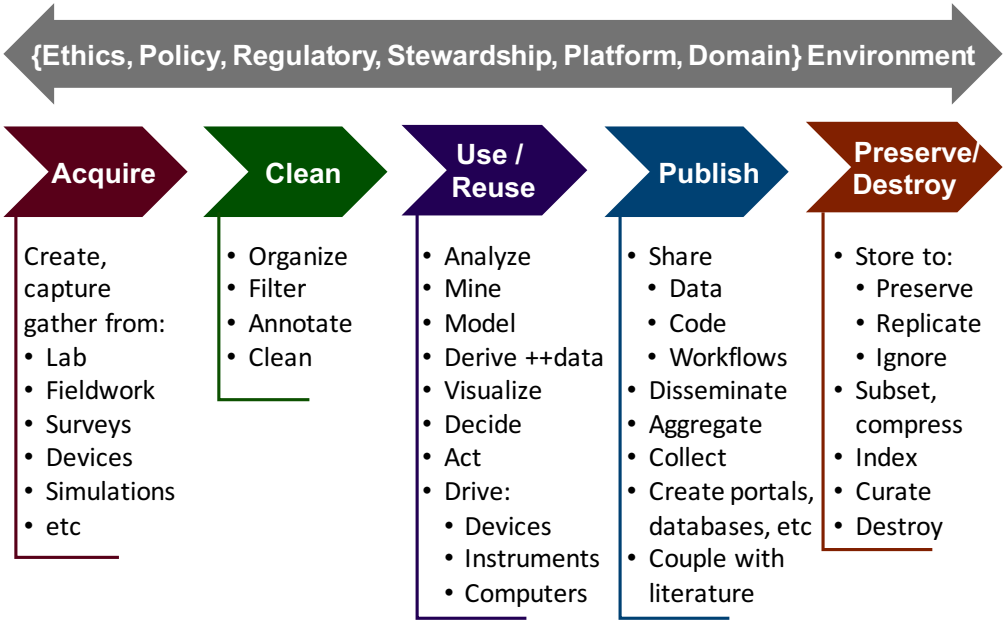

FIGURE 1: The Data Life Cycle and Surrounding Data Ecosystem

The *Data Life Cycle* is critical to understanding the opportunities and challenges of making the most of digital data. **Figure 1** shows a simplified diagram with essential components of the data life cycle. Data is *acquired* from some source (measured, observed, generated); *cleaned* and edited to remove the outliers inevitable in real-world measurement scenarios and to render it suitable for subsequent analysis; *used* (or reused) via some analysis leading to insight, action, or decision; *published* or disseminated in some way so the community at large is made aware of the data and its outcome(s); *preserved* (or not) so that others can revisit and reuse this data now or in the future. Surrounding this overall pipeline is a broader *environment* of concerns: *stewardship* to maximize the quality of the data and promote effective use, *ethics* issues that touch on proper or improper actions with these data; *policy* and *regulatory* constraints that impose legal limitations on these data; *platform* and infrastructure issues that affect technically how researchers can work with data; and *domain* and disciplinary needs specific to the application communities that create, operate, and use the data from these pipelines.

As an example of the data life cycle, consider data representing experimental outputs of the Large Hadron Collider (LHC), an instrument that is tremendously important to the physics community and supported by researchers and nations world-wide, including the U.S. LHC experiments collide particles to test the predictions of various theories of particle physics and high energy physics. In 2012, data on LHC experiments provided strong evidence for the Higgs Boson, supporting the veracity of the Standard Model of Physics.

The life cycle of LHC data is fascinating. At "birth," data represents the results of collisions within the instrument that spans a 17-mile tunnel on the France-Switzerland border. Most of the data generated is technically "uninteresting" and disposed of, but a tremendous amount of "interesting" data remains to be analyzed and preserved [Estimates are that by 2040, there will be between 10 and 100 exabytes (billion trillion bytes) of "interesting" data produced from the LHC]. Undisposed LHC data is annotated, prepared for preservation, and archived at more than a dozen Tier 0 and 1 sites. This data is published/disseminated to the community for analysis and use at over 100 Tier 2 sites. Because of critical attention to stewardship, use, and dissemination of LHC data throughout its life cycle, physicists were able to provide convincing evidence for the Higgs Boson, a scientific discovery that earned the Nobel Prize in 2013[8] and was *Science* magazine's "Breakthrough of the Year" in 2012[9].

An important part of the LHC data ecosystem is the development of stewardship, dissemination, and use protocols as well an economic model that supports the data and its infrastructure. Without this ecosystem, community agreements about how the data is organized, and political and economic support, LHC data would not meet its potential to transform our knowledge of physics and make the most of the tremendous investment in the physical instruments and facilities.

The data life cycle diagram and the LHC example suggest a seamless set of actions and transformations on data, but today's reality in many communities and disciplines is that many of these steps are deeply siloed. Domain scientists focus on generating data and using data. Computer scientists often focus on platform issues, including mining, organizing, modeling and visualizing, and the potential for eliciting meaning from analysis of the data through machine learning (ML) and other approaches. The physical processes of acquisition and instrument control are often the focus of engineering (i.e., data as "dirty signals" or as control inputs for other equipment). Statisticians may focus on the mathematics of models for representation and inference. Information scientists and library scientists may focus on stewardship and preservation of data and the "backend" of the pipeline, after acquisition/decision/action, in the realm of publishing, archiving, and curation.

There is a significant opportunity for NSF CISE to *bridge gaps* in the development of effective life cycles for valuable data within CISE, among the Computer Science, Information Science, domain, and Physical Engineering communities, for a start. There is also an opportunity to bridge gaps among ML, data analytics, and synergistic partner disciplines such as statistics and operations research. Individual NSF programs already address different aspects of the problem, e.g., BIGDATA, Transdisciplinary Research In Principles Of Data Science (TRIPODS), Computational and Data-Enabled Science and Engineering (CDS&E), DIBBS, DataNet, etc. The new opportunity is to focus on the full data life cycle (Figure 1) *across* a set of interlinked NSF programs. We describe target areas that provide a focus for these opportunities in more detail in the next sections.

III. TARGET AREAS OF A DATA SCIENCE AGENDA

Data is, by its very nature, cross-cutting. At the most basic level, data is bits that must be stored, versioned, indexed, searched, and potentially preserved. Data's value is in how they are interpreted and used. The improvement of algorithms, tools, and platforms for dealing with data is the focus of Data Science research. The development of next-generation data scientists as well as a data-literate workforce that can utilize data-driven techniques to increase innovation and success is the focus of

[8] The Nobel Prize in Physics 2013. Retrieved from http://www.nobelprize.org/nobel_prizes/physics/laureates/2013/
[9] Cho, A. (2012). The Discovery of the Higgs Boson. *Science (338)*6114, 1524-1525. doi: 10.1126/science.338.6114.1524

Data Science education and training. Critical for the successful growth and development of both research and education in Data Science are adequate platforms and infrastructure to support the data needed for successful efforts. We address research, education, and infrastructure needs of Data Science below.

A. The Need for a National Data Science Research Agenda

Almost every stage of the data life cycle (Figure 1) presents research opportunities. Certainly digital data can be viewed as an end used to provide answers directly (supporting infrastructure must enable data to be stewarded better/faster, indexed, queried, etc.). But researchers also have tremendous opportunities to adapt the data to provide value beyond simple questions/answers. Emerging tools and techniques are helping to

- **Extract** structured information from unstructured sources (dark data);
- **Clean** and curate the data so they can be used in a novel analysis (adding semantics to the "columns" and links among different parts of the data);
- **Protect** personal sensitive information while still allowing statistical conclusions to be drawn about a relevant population (data privacy);
- **Relate** different datasets together creating geometrically more value;
- **Train** embodied-intelligence systems (e.g., artificial intelligence) and other decision-oriented artifacts;
- **Evolve** such smart systems over time with the input of new information; and
- **Act** on outputs of vast data pipelines, in ways that affect individuals and society itself.

The overarching area of opportunity for a national Data Science agenda is to bridge the gaps in the development of effective life cycles for valuable data within CISE, and among the computer science, information science, statistics, and the science and engineering communities mentioned in the previous section, as well as maximizing our capability for data use and data-driven insights. Said differently, a "business as usual" agenda is likely to strengthen individual technologies underpinning discrete steps in the data life cycle but is unlikely to nurture broader breakthroughs or paradigm shifts that cut across existing disciplinary silos. It is an essential and defining attribute of big data that it can *connect* previously disparate disciplines, communities, and users, and it can serve as the vehicle to deliver *convergence* among disciplinary areas.

Thus, it seems vital to encourage a broader and more holistic view of data as an *integrating* research opportunity across the sciences, engineering, and the full range of application domains. The opportunity is to invest in data – in particular, the full data life cycle and surrounding environment – as *a central outcome itself*, and not as a side effect or intermediate step to another desirable outcome. For example, NSF might target emerging areas of fundamental science or applied engineering and use any emerging big data opportunities as the organizing principle for a set of future funding opportunities.

In addition to guiding the development of Data Science as a core component of CISE, Data Science must also evolve to address the needs of domains outside of CISE. NSF has a unique opportunity to advance Data Science, both with respect to applying data-driven strategies to individual domain research and with respect to cross-domain research opportunities.

A second opportunity is in what might be called "embodied intelligence" scenarios that big data is newly enabling. Recent breakthroughs in a range of fundamental Artificial Intelligence (AI) technologies, widely referred to as "deep learning," have made it possible to create sophisticated

software artifacts that "act intelligently." The key innovations are in mathematical pattern recognition techniques that can take inputs from millions of training examples of correct responses, to create software systems (and soon, likely hardware systems) that better recognize images, decode human speech, discover critical patterns in legal or business documents, and the like. As engineered artifacts, these AI systems are embodied as complex mathematical formulae that are customized to purpose (i.e., trained) by a truly astounding volume of numerical parameters (e.g., 10,000,000 for a decent image classification system today). Recent Federal government research and development (R&D) strategic plans for big data and AI have referred to such opportunities.[10]

These trained decision-oriented models are becoming core components in a range of novel software solutions to complex problems, creating cross-disciplinary challenges. For example, what does it mean for such a component to be "correct" when it is perhaps only 70 percent accurate? What is the life cycle for the data used to train and update these models? What are the policy implications for embodied intelligent agents trained on such data that "act badly" in the real world, e.g., that are blamed for an autonomous vehicle that crashes, or result in the suspension of a customer's account based on an incorrect automatic inference? Software Engineering, as a discipline, is challenged by such imprecision and with versioning and testing of the enormous data components (gigabyte-to-terabyte scale training data) for these systems. Existing notions of model verification/validation seem woefully insufficient. And the policy, stewardship, and curation questions are largely un-posed and unanswered.

A third opportunity is to address the growing gap between commercial and academic research practices for data systems at the edge of the state-of-the-art. Much has been made of the recent "reverse migration" of strong academic researchers into data-rich enterprises such as Google, Microsoft, Facebook, etc. While this is likely good for the national economy in the near-term, this is worrisome for the future of discovery-based open research on big data and the academic sector's capacity for discovery. One reason for the "brain drain" from the research community is the scarcity of datasets and adequate infrastructure environments in academia that support Data Science at scale. When the best infrastructure environment for cutting-edge research is consistently in the private sector, the opportunity for innovation in the public sector deteriorates. NSF development of strategic and committed public-private partnerships that build adequate and representative at-scale infrastructure in the academic community can unfetter the capacity for innovation in the research community and ultimately support the private sector through development of a more sophisticated, educated, and better-trained workforce. There may be opportunities for more collaboration with industry—it seems in industry's own interests to support such activities in academia—to further the research agenda, as well as training and education.

[10] See https://catalog.data.gov/dataset/the-federal-big-data-research-and-development-strategic-plan and https://www.whitehouse.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/national_ai_rd_strategic_plan.pdf.

Recommendations for a National Data Science Research Agenda

R.1. CREATE DATA SCIENCE RESEARCH CENTERS. Focus/re-focus a set of subsequent calls for Science & Technology Centers (STCs) and Engineering Research Centers (ERCs) as *Data Research Centers*.

> *Utilize key Data Science challenges as the thematic focus of the Centers. Encourage proposers to take a comprehensive approach that uses the full data life cycle as a central organizing principal and value proposition in addressing these challenges.*

R.2. INVEST IN RESEARCH INTO DATA SCIENCE INFRASTRUCTURE THAT FURTHERS EFFECTIVE DATA SHARING, DATA USE, AND LIFE CYCLE MANAGEMENT. Develop programs that focus attention on critical problems (privacy, inference, provenance, etc.) that remain obstacles to the use of data at scale.

> *Research outcomes should ultimately be translatable to infrastructure that enable access to data in ways that: (i) preserve privacy and other commitments made when collecting the data; (ii) enable researchers to make "unbiased" inferences, or understand potential biases in the data and other data use challenges; (iii) support reproducibility; (iv) support access, provenance, sustainability, and other life cycle challenges; and (v) support research into new hardware/software infrastructures needed to support Data Science research.*

R.3. SUPPORT RESEARCH INTO EFFECTIVE REPRODUCIBILITY. Develop research programs that support computational reproducibility and computationally-enabled discovery, as well as cyberinfrastructure that supports reproducibility.

> *Potential research efforts may focus on, for example, mechanisms to extend validation, verification and uncertainty quantification to include reproducibility; software standards (creation, test, curation, etc.); tools for sharing and verifying queries on confidential data; tools to understand links between decision-oriented models and their training data (e.g., emerging artificial intelligence models); etc.*

R.4. FUND RESEARCH INTO MODELS THAT UNDERLIE EVIDENCE-BASED DATA POLICY AND DECISION MAKING. Invest in the development of models that can be used to support data-related policy, regulation, and strategic investment.

> *For example, investments in open access efforts should be supported by models that link data value, data cyberinfrastructure, and data investments. This would provide evidence for transparent policies and processes for exploring, for example, which digital artifacts might require additional resources to be invested in preservation, and how to make such decisions at the levels of the agency, enterprise, and institution.*

R.5. EXPAND FUNDING INTO DEEP LEARNING, SMART ENVIRONMENTS, AND OTHER ARTIFICIAL INTELLIGENCE-EMPOWERED AREAS AND THEIR USE IN DATA-DRIVEN APPLICATIONS. Continue to invest in both the foundations of "embodied intelligence" research and in their use by applications and within domains.

> *Encourage research that links to new domain applications, new cyberinfrastructure, and new curricula for Data Science.*

B. The Need for a National Data Science Education and Training Agenda

It is clear at higher education institutions across the country that Data Science is a critical skill for both 21st century research and a 21st century workforce. In higher education, Data Science curricula have two audiences: new professionals in the field of Data Science, and scientists and professionals who need Data Science skills to contribute to other fields. Data Science curricula in higher education often focus on both groups, in the same way that curricula in Computer Science departments both educate computer science students and provide training in computer skills to students from other disciplines to promote computer literacy.

At present, there is no single model of which department, school, or cross-unit collaboration within higher education institutions should have the responsibility for Data Science education and training. Successful Data Science programs are being sited in departments and schools of Computer Science, Information Science, Management, and others. Many of the most successful programs in Data Science, particularly at the undergraduate level, represent university-wide coalitions frequently sponsored by interdisciplinary institutes, rather than being centered within particular departments/schools. In short, there is no common agreement as to "where" Data Science should live, but there is much interesting experimentation at this point. Hence, we focus on trends in Data Science education and training in the following sections.

Data Science Education

In terms of "what" is a Data Science curriculum, trends are emerging from programs being put together at institutions such as Columbia University, those funded by a program co-sponsored by the Moore Foundation and Sloan Foundation (at University of Washington, UC Berkeley, and NYU Polytech), and others. In general, data scientists are expected to be able to analyze large datasets using statistical techniques, so statistics and modeling are typically part of required coursework. They must be able to find meaning in unstructured data, so classes on programming, data mining and machine learning are also usually part of the core. They must be able to communicate their findings effectively, so courses on visualization are commonly offered, at least as an elective. Due to the misuse of data and incorrect conclusions drawn from data, ethics is also becoming a "must-have" in any responsible curriculum.

Other courses that appear either in the core or as electives in various programs include research design, databases, algorithms, parallel computing, and cloud computing, all of which reflect skills that an employer might expect from a data scientist. Many programs also require a capstone project that gives students experience in working through real world problems in teams within a particular domain. Interesting Data Science options are also provided in Data Science courses that are starting to be a staple of quality on-line programs.

A strong Data Science curriculum requires faculty with appropriate expertise and engagement with the field. The pull of faculty with expertise in Data Science and related fields away from academia and towards industry creates a challenge for educational institutions in mounting these programs. It also presents a potential challenge to the development of Data Science as a formal discipline.

To combat this trend, in 2013, the Moore and Sloan Foundations created a joint $38 million project to fund initiatives to create "Data Science Environments"[11] addressing challenges in academic careers, education and training, tools and software, reproducibility and open science, physical and

[11] See http://msdse.org/

intellectual space, and Data Science studies. The three institutions funded were the Berkeley Institute for Data Science, the New York University Center for Data Science, and the University of Washington eScience Institute. Not surprisingly, these institutions are developing strong educational programs in Data Science and are being encouraged to lead in this space by sharing their results and educational materials with others. Programs such as this are also likely to attract strong faculty and students, and promote Data Science research and education within the academic sector. The Moore-Sloan funding has been transformational in this respect.

The Moore-Sloane Data Science Environments have had a very positive impact, driving vanguard institutions to think deeply about creating a sustainable Data Science ecosystem within academia. This ecosystem is essential for education. Key elements of the ecosystem include the ability to recruit and hire faculty and research scientists with Data Science expertise; the importance of creating career paths for research data scientists within academia; the development of curricula that underlies current and future innovation within Data Science and in data-enabled fields, the need for supporting infrastructure for Data Science research, etc. Additional and alternative models for this ecosystem will also likely emerge if more funding of this type became available through programs at NSF. NSF would also have the opportunity to expand beyond the natural sciences (which were targeted in the Moore-Sloan funding) to other areas across the full spectrum of science and engineering supported by NSF. Further, NSF could develop programs to evaluate the effectiveness of different methods and approaches for the overall architecture of Data Science programs/curricula and for the teaching of core Data Science topics.

Currently NSF funding, in particular through the Integrative Graduate Education and Research Traineeship (IGERT)[12] and NSF Research Traineeship (NRT)[13] programs, is also impacting the development of PhD programs related to Data Science, although more work needs to be done to increase the number of projects proposed and/or funded in Data Science. Additional mechanisms should be developed to make sure that such programs can be shared with or transferred to other institutions developing programs in Data Science. In addition, a number of workshops, some sponsored by NSF, have been held to share ideas on curricular issues for Data Science. For example in the mathematical science community, the National Academies of Sciences, Engineering, and Medicine Committee on Applied and Theoretical Statistics (CATS) is actively discussing what students should be learning through their "Roundtable on Data Science Post-Secondary Education" meetings.[14]

At this time of exciting experimentation with various instantiations of Data Science curricula and placement to promote Data Science education, it is important for NSF to support and nurture a variety of approaches to develop new generations of data scientists. The NSF-sponsored National Academies of Sciences, Engineering, and Medicine study on "Envisioning the Data Science Discipline: The Undergraduate Perspective"[15] is a good first step.

Data Science Training

We distinguish here between "education," which is focused on creating core data scientist professionals, vs. "training," which exposes domain specialists to essential data skills necessary for them to advance their discipline. Data skills are needed across fields, from Geoscience to Social Science, and broadly throughout every sector. They are also needed by the consumers of the

[12] See https://www.nsf.gov/crssprgm/igert/intro.jsp
[13] See https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505015
[14] Also see National Research Council. (2013). *Frontiers in Massive Data Analysis*. The National Academies Press. doi: 10.17226/18374
[15] See http://sites.nationalacademies.org/CSTB/CurrentProjects/CSTB_175246

conclusions of Data Science (e.g., business management, law, government), so that they can accurately understand and correctly interpret the results of a study or a query.

Training in the nuts and bolts of dealing with data is critical for a data-savvy workforce. Training in appropriate software engineering skills should be provided to all students who will be using data-driven computational methods in research. Training in the use of version control and repositories for data and software should be taught and expected for all computational researchers. Training in best practices for digital scholarship should be incorporated into required curricula, including discussions of ethics. Key scientific societies or communities should consider developing standard short courses or training courses to help develop the next-generation workforce.

Data Science training should enable domain scientists and data-enabled professionals to operate effectively in the *broader* data environment. In addition to technical skills that enable professionals to apply Data Science to their efforts, it is important that Data Science training helps develop the ability to utilize data effectively and appropriately within a societal framework. This means that Data Science education and training should include a focus on policy, ethics, and privacy. In particular, many institutions are coupling courses in technical Data Science skills with courses that describe the complex environment in which data exist. This environment includes the need for considerations of

- **Privacy**: protecting human subjects in data from the risk of discoverability when linking datasets[16];
- **Agency**: human subjects may wish, or even direct, their data be shared across multiple studies or even publicly, but, at present, there are few vehicles for this (institutional review boards, etc.);
- **Compliance**: many fields using new data sources may not be aware of restrictions that come with regulations such as the Health Insurance Portability and Accountability Act of 1996 (HIPAA), the Family Educational Rights and Privacy Act (FERPA), etc;
- **Responsible research:** the conflict between doing reproducible research in the sense of making code/data/workflows openly available with the publication and the fact that this is not currently rewarded (although this is slowly changing; credit is still given for the publication, whether or not it is computationally reproducible and additional artifacts have been made useful and available);
- **Ethical statistical analysis**: applying statistical techniques as they were intended and fully reporting all statistical analysis. For example, not handpicking hypothesis test results, nor omitting negative results, training and testing models reliably, etc. Research results will be unlikely to reproduce in new samples if generated in such ways;
- **Ethical re-use**: citation of code and data use in the references section; and
- **Collaborations with third parties**: a potential conflict if, say, an industry or other partner does not want the project's code or data revealed, yet the university researcher would like to, or is required to, disseminate reproducible research.

As mentioned before, NSF has activities underway in the form of the sponsored workshops and the TRIPODS program that emphasize fundamental principles of Data Science as well as the accompanying pedagogy and education agenda. One venue for development of educational curricula and training for broad Data Science skills is within the BD Hubs program and other community-building projects. Within the context of the BD Hubs program goal to "stimulate regional and grassroots partnerships focused on big data," the Hubs are excellent environments in

[16] See, for example, Lane, J. et al. (2014). *Privacy, Big Data, and the Public Good: Frameworks for Engagement*. Cambridge University Press. Available at http://www.dataprivacybook.org/

which to pilot efforts to educate and train the big data workforce in individual regions, including data scientists, business managers, students, and end users.

Recommendations for a National Data Science Education & Training Agenda

E.1. SUPPORT THE DESIGN AND DEVELOPMENT OF DATA SCIENCE PEDAGOGY AND CURRICULA. Make data a central focus for large cross-disciplinary education research centers, courses, complete curricula, and modes of pedagogy.

> *Include curriculum foci on Data Science research areas, social issues around data stewardship and use, and data life cycle components. Create opportunities across NSF directorates and offices to collaborate on Data Science education [e.g., Mathematical and Physical Sciences; Social, Behavior, and Economic Sciences (SBE), Biological Sciences, and Geosciences]. As two examples, NSF's Critical Techniques, Technologies and Methodologies for Advancing Foundations and Applications of Big Data Sciences and Engineering (BIGDATA) program might be naturally evolved or extended to include such activities, and NSF's Big Data Regional Innovation Hubs (BD Hubs) program could be additionally resourced to focus on effective workforce education and training as part of its core mission to develop big data sources, techniques, and shared infrastructure.*
>
> *In addition, support not only Science, Technology, Engineering, and Mathematics (STEM), but also non-STEM education opportunities in Data Science. For example, consider support for educational research on infusing statistical and machine learning "thinking" into business, government, and management programs, so that the results of big data can be effectively used by business leaders to make decisions.*

E.2. TARGET EXISTING OR NEW PROGRAMS TO DEVELOP DATA SCIENCE CURRICULA AT EPSCOR AND MINORITY-SERVING INSTITUTIONS. Include curricula and training programs that assist students at participating institutions to be competitive for Data Science internships in the private sector and fellowships in the academic sector.

> *Focus and support for institutions at all tiers is critical to develop the best and the brightest Data Science workforce and to make Data Science an inclusive area for all.*

E.3. ENCOURAGE "DATA INCUBATOR PROGRAMS" OR OTHER PRIVATE/PUBLIC PARTNERSHIPS THAT PROVIDE STUDENTS/FACULTY ACCESS AND OPPORTUNITIES FOR ENGAGEMENT WITH FACULTY/INDUSTRY/NON-PROFITS WITH REAL PROBLEMS. Develop programs that train a sophisticated Data Science workforce that is more prepared to address Data Science challenges at the frontier of innovation.

> *NSF's Industry/University Cooperative Research Centers Program (I/UCRC) is a start in this direction, and could be expanded, e.g., via the NSF Big Data Regional Innovation Hubs: Establishing Spokes to Advance Big Data Applications (BD Spokes) program.*

E.4. SUPPORT PhD AND POSTDOC FELLOWSHIPS IN DATA SCIENCE. These are important to address current and future research and workforce needs.

> *Fellowship programs could be supported through NSF Data Science programs or as an extra year in other programs with a plan for training in Data Science. For example, Data Science was an emphasis area for the first round of the NSF Research Traineeship program.*

C. The Need for Datasets and Enabling Data Infrastructure to Support Data Science Research and Education

Any innovative agenda in Data Science research and education will depend on a foundation of enabling data infrastructure and useful datasets. There is a need for research infrastructure for Data Science—including experimental data acquisition testbeds; access to a variety of datasets at scale; and access to the appropriate hardware/software for computing with data at scale—distinct from the data infrastructure needed by domain scientists and engineers for data-intensive work in their respective domains. Lack of such infrastructure and datasets will retard success. Research in Data Science needs access to sufficiently large and numerous datasets to illuminate and validate results. These datasets must be available for reproducible research and hosted by reliable infrastructure. Education and training in Data Science requires an environment where students can work on data that represents the datasets and environments that they will see in the professional arena (i.e., data that is at-scale and a stewardship infrastructure that enables that data to be a useful tool in analysis, modeling, mining, and other efforts).

In the best case, data infrastructure should support access to data for research and education that is equivalent to access to any other key utility: it must be "always on," it must be robust enough to support extensive use, and the quality must be good. Research and education in Data Science requires access to data in all forms—including "raw," messy data (as encountered in most real-world situations), as well as curated data. Learning about data curation, and studying and developing tools for automated curation, would be part of a Data Science agenda. Deployment in real-world situations would be in close collaboration with the appropriate domain science or engineering disciplines. Data Science can highlight the fact that making data part of the infrastructure comes down to responsible stewardship (who is committed to keeping it, in what formats, for how long, etc.); availability of provenance information that promotes effective use (who created the data, curated the information, used the data, and deployed solutions based on the data and how was this done); ease of access, use, and discoverability (useful solutions for how researchers find data, who can access it, and what services are needed to make it useful); and other key infrastructure components (state-of-the-art storage and migration plans, useful tools for dissemination and visualization, adequate privacy and security controls, and embedded professionals to assist with curation, intellectual property and policy concerns, reporting, etc.).

As part of the data ecosystem, it is also critical to develop clear and useful policies with respect to how organizations, institutions, and projects deal with data (what data is kept, who owns it and its byproducts, and who has access to the data or to parts of it).

Since data will become the core for research and insight for a broad set of academic disciplines, access to data in a usable form on a reasonable time scale becomes the entry point for any effective research and education agenda. NSF has an opportunity to ensure that lack of adequate data infrastructure is not a roadblock to innovative research and educational programs.

Developing and sustaining the infrastructure that ensures that research data is available to the public and accessible for re-use and reproducibility requires stable economic models. While there is much support for the development of tools, technologies, building blocks, and data commons approaches, there are few Federal programs that directly address the resource challenges for data stewardship and provide needed help for libraries, domain repositories, and other stewardship environments to become self-sustaining and address public access needs.

While NSF cannot take on the entire responsibility for stewardship of sponsored research data and its infrastructure, neither should the Foundation shy away from providing seed or transition funding for institutions and organizations to develop sustainable stewardship options for the national scientific community. In particular, NSF should work with institutional libraries and domain-specific repositories to develop and pilot sustainable data stewardship models needed for data-driven research through strategic programs, guidance, and cross-agency and public-private partnerships.

Recommendations for National Data Infrastructure that Supports Data Science

I.1. SUPPORT THE CREATION, ACQUISITION, AND PUBLIC DEPLOYMENT OF A BROAD PORTFOLIO OF REALISTIC, STATE-OF-THE-ART, AT-SCALE DATASETS FOR ACADEMIC RESEARCHERS. Create collections of educational datasets and materials and make them available via publicly accessible repositories, libraries, and stewardship environments. Link them to the publications they support.

> *Creation of this fundamental infrastructure will drive increased inter-sector relationships (particularly with the private sector) to acquire and deploy datasets useful for scholarly research and Data Science educational programs.*

I.2. DEVELOP AND DEPLOY DATA SCIENCE INFRASTRUCTURE NEEDED TO SUPPORT CUTTING-EDGE RESEARCH AND EDUCATION. Invest in both national and institutional infrastructure to support emerging Data Science research and education programs.

> *Focus on low-barrier-to-access, representative, and sustainable Data Science infrastructure that supports research efforts and coursework. Initiate strategic and committed public-private partnerships that can help build representative at-scale infrastructure in the academic community for Data Science research and education.*

I.3. CREATE A NATIONAL EFFORT TO WORK WITH INSTITUTIONAL LIBRARIES AND DOMAIN REPOSITORIES TO DEVELOP SUSTAINBLE MODELS OF DATA STEWARDSHIP AND PRESERVATION OF VALUED SPONSORED RESEARCH DATA. Work with the community to develop and ramp up the use of sustainable institutional economic models that enable participating organizations to provide open access research data stewardship options to the community over the long term.

> *Evaluate outcomes thus far from NSF programs like Sustainable Digital Data Preservation and Access Network Partners (DataNet) and Data Infrastructure Building Blocks (DIBBS) and use lessons learned to inform an effective program.*

I.4. ESTABLISH BEST PRACTICE GUIDELINES FOR DETERMINING THE BALANCE OF DATA SCIENCE INVESTMENT THAT SHOULD BE FOCUSED ON RESEARCH VS. ENABLING INFRASTRUCTURE. Look to other organizations, institutions, and sectors to establish a rough guideline about the distribution of investment to maximize the effectiveness of research and education efforts.

> *To capitalize on infrastructure investments, encourage proposers to utilize existing useful data infrastructure when possible and to clearly articulate the value added for new infrastructure efforts in proposals.*

I.5. SUPPORT EFFORTS TO GATHER KEY INFORMATION ABOUT THE SCALE AND STATE OF SPONSORED RESEARCH DATA ACCESS AND STEWARDSHIP TO INFORM EVIDENCE-DRIVEN INVESTMENT AND POLICY. Gather statistics for sponsored research that describe the amount, characteristics, and stewardship of the data and allow linking between data and scholarly publication, perhaps through vehicles such as Fastlane or Research.gov.

> *Relevant characteristics may include amount and type of data generated, where the data is hosted, whether the data is available to a broader community, what publications the data is associated with, etc. Make this information available through the National Science Board Science and Engineering Indicators or a similar publication.*

IV. ON THE HORIZON: AREAS THAT WILL EXPAND DATA SCIENCE

Beyond current and near-term needs to grow Data Science and Data Scientists to meet the needs of modern research, education, and training, it is important to recognize that Data Science will be fundamental to realize the potential of future data-driven scenarios. As systems grow "smarter" and take on more autonomous and decision-making capabilities, society will increasingly be faced not just with technical challenges but also with the social challenges of governance, ethics, policy, privacy, and other issues that will render data-driven systems useful, effective, and productive, rather than intrusive, limiting, and destructive. And, as fundamental computing platforms change, there are large opportunities to reimagine the entire hardware/software enterprise in the light of future data needs.

NSF can play a role in ensuring that scientists are prepared to deal with future scenarios by encouraging the initial research and efforts that lay the groundwork for well-functioning systems, useful policy and protections, and effective governance of data-driven environments. We briefly describe three opportunities here.

Data-Enabled Policy and Evidence-Based Decision Making

At this juncture, policymakers are coming under increasing pressure to utilize scientific knowledge to underlie decisions. According to a 2014 Pew-MacArthur Results First Initiative report[17], "Recognition is growing that policymakers can achieve substantially better results by using rigorous evidence." For example, the Office of Strategic Environment Management within the Environmental Protection Agency advises on evidence-based decision making and program implementation, and the Office of Management and Budget issued a memorandum in 2013[18] exhorting Federal agencies to "improve program performance by applying existing evidence about what works, generating new knowledge, and using experimentation and innovation to test new approaches to program delivery." Most recently, the House of Representatives passed the Evidence-Based Policymaking Commission Act of 2016[19], suggesting that Congress is taking notice of the value of Data Science findings. And the White House Office of Science and Technology Policy's National Science and Technology Council established a Data Science Interagency Working Group, focused on "data-driven decision making by advancing adoption of data centric technologies, techniques, best practices, leveraging Federal R&D efforts. It will identify the most pressing challenges in innovation using data science…"[20]

What is involved in representing the methodologies and data of researchers in creating reliable and reproducible results? Workflow system capture becomes imperative for tracking research steps, code is needed to encapsulate research protocols and data analyses, and project datasets must be retained, annotated, and curated. Beyond creating the "recipe" for the research process, the process of interpreting research results itself is challenging. For example, what defines a "trustable" result? Should the public have a right to inspect the processes that led to the decision, or in the case of research, the scientific conclusion? What information could be provided to help

[17] The Pew-McArthur Results First Initiative. (Nov 2014). *Evidenced-based Policymaking: A Guide for Effective Government.* Retrieved from http://www.pewtrusts.org/~/media/assets/2014/11/evidencebasedpolicymakingaguideforeffectivegovernment.pdf
[18] Office of Management and Budget. (2013). *M-13-17 Memorandum to the Heads of Departments and Agencies*. Retrieved from https://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-17.pdf
[19] 114th Congress. (2015-2016). *H.R.1831 Evidence-Based Policymaking Commission Act of 2016.*
[20] Kalil, T. (2016). *Charter of the Data Science Interagency Working Group Committee of Technology National Science and Technology Council*. Retrieved from https://www.whitehouse.gov/sites/default/files/microsites/ostp/NSTC/Data%20Science%20IWG%20Charter-%20Signed%20(1).pdf

non-specialists and non-scientists understand decision processes and the scientific research problems?

These questions are important for the development of data-driven, evidence-based policy and require research and experimentation for better approaches for capturing the research workflow and products so they can be examined and reproduced. NSF can foster the ample discussions needed to address such questions through workshops, community discussions, data literacy curricula, as well as through specific research projects and efforts. We note that these issues clearly transcend CISE and are Foundation-wide concerns. Our goal in this discussion is to underscore the need for a strong connection from the technical side of the *Data-Knowledge-Action* cycle to critical policy outcomes.

Data and the Internet of Things

As data increasingly drives research, education, commerce, and modern life, and as information systems become more ubiquitous, society is rapidly moving towards increasingly integrated cyber-physical-biological systems. The *Internet of Things* (IoT) refers to an ecosystem in which systems, devices, people, and organizations are highly interconnected and that supports targeted analysis, autonomous decision making, and enhanced capabilities.

Data Science will be fundamental to creating, exploring, and supporting the data-driven and autonomous systems of IoT. Long-term investment by NSF in Data Science that will drive IoT can lay the groundwork for data-driven systems that are safer, more reliable, more secure, and "smarter" to support a broad spectrum of anticipated and unanticipated uses.

Along with technological innovation within Data Science, society will also see the need for increasing structures that define the social interactions of a world where the actors are both humans and machines. Current challenges around the ethical use of technology will become exacerbated when decisions are made autonomously by machines as well as humans (e.g., how does a self-driving car choose between two bad options?). The development of "artificial ethics" systems that complement AI systems will become increasingly important in this scenario.

What is NSF's role? At this point, much of the initial exploration, study, and groundwork must be laid. NSF can help the research community get out in front of these problems and promote the critical initial discussions for the ethical and effective use of IoT technologies. This can be accomplished by building upon investments that CISE is already making in programs like Smart & Connected Communities (SC&C)[21] and Secure and Trustworthy Cyberspace (SaTC)[22] and working with other NSF directorates and offices to research technologies and technical infrastructure that will underlie effective use of IoT, as well as continuing to support studies, workshops, and community discussions about privacy, rights, governance, and other "social infrastructure" needed to realize the potential of IoT.

Data and the End of Moore's Law

The rise of big data, and the simultaneous demise of Moore's Law scaling of semiconductors, creates a surprising set of opportunities. The end of the shrinking of transistors is nearing, limiting the essence of scaling. Transistors are today measured in nanometers (nm); the distance between two

[21] See https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=505364
[22] See https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504709

silicon atoms in the crystalline lattice of a modern chip is roughly 0.5 nm. Current computing platforms are shipping with semiconductors fabricated at the so-called 10nm node. The next two planned nodes are 7nm and 5nm, respectively. After which, scaling essentially *stops*. To the extent that computer architecture can be regarded as the art of transforming transistors in performance, future generations of computers will no longer be able rely upon improvements in the fundamental circuit fabric itself. Tomorrow's transistors will likely *not* be smaller, cheaper, faster, or lower power, in any intrinsic way. What is left is to *organize* them in ways that are *smarter*. More specifically, computing platforms must be architected to exploit opportunities intrinsic to the applications they are intended to run.

Already, we are seeing the emergence of such hardware platforms in data-centric applications. The Microsoft Bing search engine has been enhanced with reconfigurable field-programmable gate array (FPGA) hardware, delivering better performance per watt. It is part of the larger Microsoft Catapult project that seeks to add FPGA fabric to customer cloud services.[23] Intel recently completed its largest-ever corporate acquisition, buying Altera, the world's second-largest reconfigurable logic chips (i.e., FPGA) supplier, with intent to deploy this technology in enterprise-level, data-center computing.[24] Deep networks are routinely trained and deployed on graphics processing units (GPUs).[25] Google has prototyped a custom silicon accelerator architecture it calls a "Tensor Processing Unit" (TPU) specifically to run tensor-centric machine learning workloads.[26]

As data-oriented tasks become an ever-larger component of future computing workloads, there is a huge opportunity to radically reimagine fundamental hardware architecture, the hardware/software boundary, and the classical software stack that lives "above bare metal"—as also discussed by NSCI. Future databases, data mining and analytics services, large-scale storage systems, and the like, will not be able to improve on their scalability and performance if the fundamental computing platforms on which they run cannot be improved. This rise/demise scenario, or ever-more data on not-better transistors, is likely to prove to be a large and fruitful area of future research in basic computer organization.

[23] See https://www.microsoft.com/en-us/research/project/project-catapult/
[24] Clark, D. (Dec. 28, 2015). Intel Completes Acquisition of Altera. *The Wall Street Journal.* Retrieved from http://www.wsj.com/articles/intel-completes-acquisition-of-altera-1451338307
[25] For example, see https://developer.nvidia.com/deep-learning-resources
[26] Jouppi, N. (May 18, 2016). *Google Supercharges Machine Learning Tasks with TPU Custom Chip* [Web log post]. Retrieved from https://cloudplatform.googleblog.com/2016/05/Google-supercharges-machine-learning-tasks-with-custom-chip.html

Recommendations for New Data Driven Scenarios

N.1. STRENGTHEN INTERNET OF THINGS (IoT)-RELATED DATA ECOSYSTEM. Focus on data aspects of IoT in critical areas such as data security, data privacy, and smart systems, as well as infrastructure supporting governance, policy, ethics, etc.

> *Broaden efforts to include consideration of ethics and data infrastructure needed for social governance of data and data policy. Provide leadership for defining and studying IoT data governance and social issues (perhaps in collaboration with SBE or through initiation of National Academies of Sciences, Engineering, and Medicine studies); develop pilot environments that can prototype prospective policies and technologies. Leverage programs and efforts like Smart and Connected Communities, Secure and Trustworthy Cyberspace, and others, to ensure that IoT is a driver for new research discovery as well as commercial innovation.*

N.2. SUPPORT FUNDAMENTAL RESEARCH ON RADICAL HARDWARE AND SOFTWARE SYSTEM ARCHITECTURES THAT TARGET EMERGING DATA-INTENSIVE TASKS. Focus on research that promotes innovative data analysis methods for next-generation scenarios and the evolution or re-thinking of the fundamental computing platforms on which they might best be executed.

> *The impending end of Moore's Law and the rise of big data are already pushing data-oriented tasks onto GPUs, field-programmable gate arrays (FPGAs), accelerators, etc. Support fundamental research on radical hardware and software system architectures that target emerging data-intensive tasks. Leverage relevant efforts from the National Strategic Computing Initiative (NSCI).*

Summary

The National Science Foundation has an opportunity to provide leadership for Data Science over the next decade and beyond. Attention to deep efforts in research and current and future scenarios that expand the field and its impact, and broad efforts in education, training, and infrastructure are needed to help Data Science reach its potential for transforming 21st century research, education, and workforce. The CISE Advisory Committee and its Data Science Working Group are tremendously supportive of NSF's current and future focus in this area and encourages NSF to use the Recommendations herein to contribute to a visionary and high-impact agenda to realize the potential of this cross-cutting field.