NSF BDPI-2016 Workshop Report

Primary Authors:

Lisa Singh
Amol Deshpande
Wenchao Zhou

Section Authors

Srinivas Aluru
Magdalena Balazinska
Gaitam Biswas
Auroop Ganguly
Aditya Johri
Fang Liu
Michael Mahoney
Chris North
Kunle Olukotun
Aarti Singh
Adam Smith
Suresh Venkatasubramanian

1 Table of Contents

2 Executive Summary

The BIGDATA program at the National Science Foundation seeks to advance research and education in computer science, statistics, computational science, mathematics, and different domain sciences specific to the field of data science. While many goals exist, one overarching goal of big data research is to develop methodologies, algorithms, and tools that are able to provide insight for different domain problems using noisy, partial, and potentially biased data sets. To further exacerbate the problem, we need to develop ways to do this while considering data ethics, data privacy, and data provenance. This workshop brought together principal investigators that are currently being funded by the BIGDATA program at NSF, program officers from different government agencies, principal investigators from a parallel program funded in Japan, and Big Data Hub Directors. PIs shared their research through poster sessions, funding opportunities were shared by program officers who participated in two panels and the Big Data Hub Directors in a third panel, and cutting edge research was shared by various experts throughout. The highlight of the workshop was the breakout sessions and discussions associated with current challenges and future directions of research and innovation in each of the breakout session areas.

While this report describes all the components of the workshop, it focuses on the ideas shared during the different breakout sessions. Based on these ideas, the report also contains a set of recommendations in the form of high impact priorities that are important, but challenging, low hanging fruit that are fairly easy to implement, and grand challenges that will promote long term interdisciplinary research within the field of big data. Main themes that emerged throughout the report included: the need for big data ethics and privacy standards, more flexible funding mechanisms for interdisciplinary research, an available platform for large-scale research, integrated tools for all stages of the data science life cycle, educational guidelines in this new discipline, and more synergistic research ties between big data theoretical research and applied research.

3 Workshop Goals

In April 2016, Principal Investigators (PIs) from Georgetown University and University of Maryland, College Park hosted a two-day NSF sponsored workshop for PIs who have received funding from the cross-directorate BIGDATA Program. The Big Data PI workshop (BDPI-2016) was organized by Lisa Singh, Amol Deshpande, and Wenchao Zhou.

The goal of BDPI-2016 was to bring together PIs and co-PIs currently funded by the Big Data program at NSF along with selected industry and government invitees to discuss current research, identify current challenges, and discuss promising future directions. The meeting was also meant to serve as an opportunity to foster new collaborations and share accomplishments.

4 Workshop Overview

Given the interdisciplinary nature of the participants, it was important to have a program that allowed PIs to interact and learn about each other, as well as learn about different projects in this broad area of Big Data. We hoped that this would further facilitate discussion in breakout sessions. Given these needs, we designed the program to include the following:

- Short talks about different types of big data projects, e.g. smart cities, personalized medicine, climate, and big data concerns, e.g. the ethics of using big data and privacy.
- Multiple funding panels, one focused on NSF funding and another that focused on big data related funding at different agencies, including NSF, NIH, NASA, and DHS, DOT, OFR, and NIST.
- A Big Data Hub panel that described the goals of the different hubs, as well as the program as a whole.
- International speakers from Japan who shared their collaborative research.
- A poster session for PIs to share their research and see what other PIs are working on.
- Break out sessions focused on core areas of innovation in the big data space.
- Table conversation questions during meals to encourage discussion.

The full agenda with slides can be found on the website (http://workshops.cs.georgetown.edu/BDPI-2016/agenda.htm). We have also included the agenda at the end of this report.

5 Breakout Session Discussions

A central component of the workshop was the breakout sessions. When participants registered, the registration form asked them to identify up to three different emerging research areas related to big data. Using that information, the workshop organizers identified the following breakout sessions:

1. Big Data Privacy & Ethics
2. Big Data Systems
3. Data Science & Visualization
4. Forming Interdisciplinary Partnerships
5. Foundational Big Data Algorithms and Theory
6. Large-scale Inference & Learning
7. Learning Analytics & Education

The organizers then recruited one or two workshop participants to lead each breakout session. The leaders of the breakout sessions are listed next to the session name on the agenda. They are also the additional authors of this report.

Breakout leaders were given a Powerpoint Template specifying the types of information to focus on during the breakout discussions. We proposed focusing the discussion on the following areas: overarching themes in the area, recent successes, obstacles impeding more rapid progress, areas of neglect, and strategic priorities and investments that will advance innovation. The slides for each breakout session can be found on the workshop website. These slides and the presentations by the workshop discussion leaders served as the foundation for the discussion that follows.

5.1 Big Data Privacy & Ethics

5.1.1 Current Landscape

The size, variety and availability of huge data sets has increased so rapidly that awareness and understanding of privacy and ethics issues surrounding the use of big data has not kept up among the general public or even among researchers. Numerous uses of big data are ethically controversial—think, for example, of computerized sentencing scores, teacher evaluation metrics, behavioral experiments conducted on users without consent, and the unintentional release of sensitive user data through poorly anonymized data sets or poorly protected interfaces. These issues arise in part because there are no clear guidelines or consensus about how private information should be used, nor about what types of transparency and accountability to expect from data-driven systems.

With the availability of mobile data, social media data, and various open data sources, there are concerns both about surveillance and how easy it is for systems that we understand poorly—from the vast ecosystem of online advertising and data aggregation, to social media-driven news consumption, to automated systems for predicting creditworthiness, recidivism or job performance—to monitor and influence our lives. What does it mean to use big data ethically? Can data ethics be built into an algorithm? While some researchers are beginning to think about data ethics, this field is in its infancy.

Similarly, our understanding of privacy concerns is only nascent. To what extent can we develop useful and commercially viable systems that make use of sensitive information while allowing users to retain control of their data? How can we reap the benefits of massive, distributed data sets while respecting basic freedoms? What do basic freedoms—principles first formulated well before the information revolution—mean in a networked world?

A fundamental aspect of both privacy and ethics issues is the asymmetry of information between companies/large organizations and individuals. Many basic questions—transparency, accountability, privacy—are driven by the need to mitigate, or at least understand and control, this profound asymmetry.

5.1.2 Major Challenges and Obstacles

In this section we discuss the different challenges and obstacles associated with data privacy and data ethics. We group these challenges into three areas: standardization challenges, algorithmic discrimination challenges, and cultural obstacles.

Standardization Challenges: Standardization for systems is always a challenge. However, standardization for data usage may be a larger one. If companies collect data about customers, should they be forced to not sell it if it contains PII data? What standards and best practices are appropriate for making data available for internal analysis within a large company? Between government organizations? How can inferences made based on private information be made available and understandable to individuals? How can users correct erroneous records or inferences? In order for standardization to be an option, either companies must be willing to adhere to standards, or governments must be willing to enforce them through regulation. These questions are complicated by the lack of national boundaries on data. Information, especially in the cloud, has no well-defined location.

Algorithmic Challenges: While there are general algorithm design and analytic system design challenges when using big data, there are also specific concerns with regards to privacy and ethics. Even if an algorithm produces meaningful, accurate results about individuals, it may be unintentionally discriminating against different populations. For example, some features that are selected may have high correlation with race, gender, sexual orientation, and when used for decision-making, may inadvertently discriminate against certain populations. The types and number of variables that are being used for big data analytics is so large, that is unclear how best to ensure this form of discrimination is not occurring. Of course, not understanding how to properly use these different algorithms is also an issue, e.g. poor inputs lead to potential privacy and ethics issues as well. The situation is exacerbated if the data being aggregated are from public sources and there is no understanding about the accuracy of these data or of the inferences. These data are being used in a way that is different than their original intent and, therefore, may contain noise, bias, etc.

Similarly, there remain many technical challenges in the design of privacy-aware systems. For example, even for settings where well-established notions of privacy or security—such as differential privacy or secure function evaluation—apply, we still do not understand what level of accuracy is achievable or how to design scalable, practical systems that satisfy the desired guarantees.

Cultural Obstacles: Everything is recorded and never removed. As a society, we accept that as normal. However, this acceptance is a problem. It is an indication that we are willing to give away privacy for services, e.g. free email, search, recommendations, etc. If society is not outraged about unethical uses of their data or about the lack of online privacy, then getting companies, governments, and others to invest time and money into this issue becomes a challenge.

5.1.3 Strategic Priorities

While many different directions were discussed during the break out session, we focus on four that will have large impacts in this area. The first focuses on awareness and standards, while the others are research directions.

1. Support workshops at multiple levels—from regional to international—to develop guidelines for big data privacy and ethical uses of big data. These guidelines can then serve as the foundation for documents related to human-subject big data issues. They can also be used to have a more informed discussion about data privacy and data ethics within the computer science community, as well as more broadly. Educating the population about privacy risks needs to increase.

2. Promote the development of protocols and algorithms that partially encrypt data and use these partially encrypted data directly in algorithms to compute values that are also partially encrypted.

3. Introduce formal verification / auditing for data. This will allow us to better understand whether or not formal verification can be used for determining fairness of a data set, determining with learning global properties of a data set cause systemic bias.

4. Develop approaches for people to interact with their data, see it, update it, and update predictions. We need to support research that incorporates mechanisms for human intervention to correct unintended mistakes of big data analysis.

5.2 Big Data Systems

5.2.1 Current Landscape

The community has made significant progress towards providing system support for big data curation, management, and analysis. There has been a tremendous amount of effort in developing increasingly efficient, open-source big data systems, including Hadoop, Spark, Impala, Myria, Asterix, GraphLab among others. The progressively mature cloud infrastructures and services have made cloud increasingly easy to use and increasingly affordable, which encourages system developers to deploy their systems as a cloud service. Currently, there are provisions of both data management and machine learning systems in the cloud (such as the offerings from Amazon, Azure, and Google).

Systems that perform information extraction and organization across text, images, and others to create knowledge bases in vertical domains: e.g., DeepDive, Google's Knowledge Base and Knowledge Vault, IBM, etc. More datasets are becoming public. Examples include Amazon public datasets and the open data movement in government.

There have also been efforts that allow ordinary users to more easily perform "big-data" analytics and calculation tasks on diverse computer environments. Systems that start from high-level DSLs and compile down to efficient plans specialized for different architectures: e.g., Delite, Spark NVL, Tensorflow. These systems, for example, have a simple switch to run on GPUs vs CPUs vs clusters.

5.2.2 Major Challenges and Obstacles

In this section we discuss the challenges and obstacles in developing big data systems. We group these challenges into four areas:

Compatibility and Unification Challenges: Big data is a mix of traditional analytics (SQL), complex ML models (deep learning, collaborative filtering, SVM) and linear algebra. There is a need to unify abstractions across different domains, such that tools and systems developed in each domain can be integrated with minimal effort. The breakout discussion also identified a mismatch of tools across scales, for example analysis that is initially written on a laptop, needs a significant rewrite to port to a cluster environment. Future hardware is increasingly heterogeneous, but there are still no abstractions for shielding complexity.

Curation and Preservation Challenges: The exponential data growth rate poses challenges for data garbage collection and curation for long-term preservation. System admins need to be able to assign meta-tags to data, which could then be used to decide, e.g., where data resides, or which data to retain, or which data should be cited, or who owns/claims/wants the data. A systematic approach needs to be developed to automate the process of data curation and preservation.

Configuration Challenges: Data science requires non-trivial understanding and expertise in data management and machine learning, for example, it requires an expert to answer questions such as how to choose ML algorithms, based on statistics of given analysis task and data set. These type of configuration and performance debugging is often times very challenging and tedious. Currently, this configuration process still largely depends on human admins; the systems are not autonomic enough.

Storage Obstacles: Storage is the slowest component in a computer system, and is often the bottleneck of system performance. Research in persistent memory is promising, but the technology is still too costly to be widely adopted. Big data is by definition out of core; we need to develop much better multi-tier systems that progressively move data across the memory/storage hierarchy.

5.2.3 Strategic Priorities

While many different directions were discussed during the break out session, we focus on three research priorities that will have large impacts in this area:

1. End-to-end data science pipelines: To achieve this, priorities should be given to innovations across the stack (OS, networking, PL, compilers, hardware, database system). For example, support research that develops technology for supporting low-latency analytics and tracking provenance and metadata. To support complex analysis pipelines, it is important to further advance research in bridging big data and machine learning.

2. Data acquisition and cleaning: Data cleaning and integration remain important topics to fund, for example, in the fields such as processing uncertain or probabilistic data, managing data with error bars, capturing data distributions, and supporting pipelines that explicitly need to process data with errors.

3. Reproducibility, long-term preservation, and sharing: There is a growing need to fund work in making datasets available and curated such that users can leverage those datasets effectively. As data grows at an exploding rate, we need to have

good models to preserve the data and the analysis: for example, keeping versions of software and analysis pipelines might not be enough; we need to retain the ability to run old code. In addition, workflows can get increasingly complex; maintaining provenance and supporting related queries would be an efficient approach to understand data dependencies and causalities.

5.3 Data Science and Visualization

5.3.1 Current Landscape

Data science is a growing area of interest. Educational programs, particularly Master's programs, are beginning across the country. They are in high demand. Computer science demand is also growing and resources in these two areas are increasing.

There is also a growing data science industry. To support this new industry, more and more libraries and tools are being developed. The growth in data science has happened in tandem with the growth in visualization and visual analytics. High performance computing has made big data visualization more feasible and allowed for in-situation visualizations to be applied in different domains. Another area of growth is virtual reality, which is not only used in gaming, but also for training in many industries, including medicine and law enforcement.

5.3.2 Major Challenges and Obstacles

This section discusses the different challenges and obstacles associated with data science and visualization. We group these challenges into five areas: data visualization representation, data science process, user intent modeling and inference, crossing the domain gap, and data science curricula.

Data and visualization representation: While methods and tools for visualizing data are continuing to emerge rapidly, the algorithms used to generate the results are still a black box. Going beyond the understanding of data through visualization, we need to find ways to use visualization to improve people's understanding of the data science process, including machine learning algorithms.

Data science process: While tools are emerging for different steps in the data science process, completing the entire process from data collection to data cleaning to data integration to machine learning to visualization is difficult for many users. Without more integrated tools, data science cannot be accessible to those without degrees in the field.

User intent modeling and inference: Cognitive scientists and human computer interaction researchers design technologies that improve the interaction between humans and algorithms. But there is still a significant gap in interactive data science between analytical methods and their users– understanding user intent and learning to infer it. As more human behavioral data becomes available, a need will grow for interpreting and inferring user intent in data science applications.

Crossing the domain gap: It is very difficult to transfer application domain knowledge to non-domain experts, such as computer scientists. It is also difficult for computer scientists to explain their methods and algorithms to domain experts. The variation in language between disciplines further exacerbates the issue. This makes collaborating, sharing techniques, and sharing data difficult.

Data science curricula: Turf wars are already beginning with regards to data science – what disciplines own the field? How can we teach students with non-traditional backgrounds? Data science courses should contain new material, but most teachers have retrofit existing computer science or statistics courses, added a little, and called it data science. Is this the right approach – should it be a mashup of topics from more disciplines? How do we create a coherent curriculum for a diverse, interdisciplinary student population?

5.3.3 Strategic Priorities

While many different directions were discussed during the breakout session, we focus on four that will have large impacts in this area. The first focuses on strategic workshops to define this new discipline, while the others are different approaches for advancing research and improving accessibility to the fields.

1. Conduct workshops with data science thought leaders and educators from different disciplines to define data science as a field. Compare curricula, share examples and develop a skeleton curriculum standard for data science. Without this, we have no way of knowing what the foundational skills are that are taught to data science students. We need to raise data science to the level of other sciences by identifying fundamental methods, open research questions, and a common language for data science specific concepts.

2. Increase NSF's focus on research programs for human-in-the-loop data science and visual analytics processes and that combine data visualization and data analytics. NSF should consider reviving programs such as FODOVA. New programs should fund the "whole stack", with emphasis on interactive methods that exploit both human cognition and automated computation. This will support research on human interaction with data and analytical processes.

3. Core to advancing the field is to increase the accessibility of fundamental data science concepts and methods. This means improving the accessibility of machine learning and visualization concepts. Educate more broadly with data

science vis-4-all and ml-4-all curricula. Conduct research programs on data science usability issues. Establish common toolkits for practitioners by connecting academia and the rapidly growing data science industry. Connect toolmakers and tool users in participatory design strategies to ensure tools are designed appropriately for users.

4. The data science process currently is slow and cumbersome. Projects that speed up the process should be supported. Areas of neglect include: complete pipelines in data science from data to extracting meaning; novel interactive environments for data science such as immersive analytics; collaborative analytics tools; interactive data cleaning and feature engineering; online data sharing and curation (accessible repositories need to be supported); and benchmarks development (such as NIST type competitions to develop standards).

5.4 Forming Interdisciplinary Partnerships

5.4.1 Current Landscape

Interdisciplinary partnerships, wherein computer scientists and domain experts work together to solve specific problems faced in an application domain, are crucial to the success of the "big data". Over the last two decades, there have been numerous instances where such partnerships have led to significant discoveries and innovations. Some examples of such partnerships, roughly ordered by their maturity, include human genome analysis, climate and weather modeling, social computing, self-driving cars, human brain initiative, and precision medicine. However, we still have a long way to go in understanding how to build and incentivize fruitful interdisciplinary partnerships. Currently, there is a lot of focus on "methods", where researchers, particularly computer science researchers, are designing new, improved, and general techniques for solving a variety of big data problems. However, the fraction of those techniques that lead to discoveries in any specific field is still quite small. There is disconnect between the sophisticated, general techniques that are being developed, and those specific questions that the domain scientists need answered.

5.4.2 Major Challenges and Obstacles

This section discusses the different challenges and obstacles associated with forming interdisciplinary partnerships. We group these challenges into three areas: lost in translation issues, lacking inter-domain communication, and lack of cross agency data/problem sharing.

Lost in Translation (Methods <--> Domains): Even when there is a good interdisciplinary collaboration, there is still a lot that is lost in translation. The computer scientists, the statisticians, or the mathematicians think in terms of methods and try to find problems to which they can apply those methods, while understanding the main problems at a shallow level. On the other hand, domain scientists don't understand computer science well enough to be able to figure out what solutions may be applicable to their specific problems. There are no real incentives for the crucial communication phase and for someone to truly double up as a cross-discipline expert; instead many researchers try to understand the other domain sufficiently to be able to claim they're doing cross-disciplinary work. As the low-hanging fruits are gradually plucked, deep expertise in both the application domain and computer science is required to make further fundamental advances.

No Inter-domain Communication: Similarly, there is very little knowledge transfer across different domains today. For example, researchers working in climate modeling and those working in public health do not interact as much, and thus the discoveries, insights or lessons learned from one field are often re-learned in another field. Benefits of such cross-domain fertilization are immense, but the barriers for achieving it are high. There are few repositories or benchmarks or taxonomies for cross-domain learning of big data applications and success/failure stories with domain nuances; further, there are few publication venues considered high impact enough for cross-domain lessons learned or for interdisciplinary work.

Lack of cross-agency data/problem sharing: Third, although there have been many efforts towards this, cross-agency coordination in terms of jointly enabling PIs to truly work on multiple disciplines is still lacking. Proposals that overlap with the mission of multiple agencies are often hard to get funded because agencies may believe that the key innovations are not aligned with their missions.

5.4.3 Strategic Priorities

While many different directions were discussed during the break out session, we focus on three research priorities that will have large impacts in this area:

1. The funding agencies should fund regular workshops run by PIs, where one (or more) PIs from methods and one (or more) PIs are from application domains. The agencies should also try to fund "seedling" (not just "eager") proposals just to encourage communication between different domains, with outcomes such as position papers and full interdisciplinary proposal development.

2. The funding agencies should develop new solicitations for funding the development of cross-domain benchmarks and repositories. They should also consider developing new solicitations that require 3-way (or more than 3-way) cross-pollination (e.g., Big Data, Public Health, and Climate; or Big Data, Education, and Social Sciences).

3. Agencies that are members of groups like NITRD should identifyi data sets and problems that cross multiple agencies and then develop mechanisms for facilitating cross-agency data sharing. Having calls specific to incentivizing their use should be developed.

5.5 Foundational Big Data Algorithms and Theory

5.5.1 Current Landscape

Foundational big data algorithms and theory encompasses a large number of cross-cutting topics across the entire spectrum of big data, making it hard to summarize all the major advances or open questions. The breakout session focused its discussion on a few major questions, but these should not be taken as an exhaustive list.

One of the major goals today is to develop and understand conceptual models of data; a given set of data can be described as points in a multi-dimensional space, or as a matrix, or as a graph, and so on, and the specific model chosen can have significant consequences on the types of techniques or algorithms that can be applied and the types of tools that can be used. Thus, designing new models of data, suitable for new and emerging application domains, is a key challenge. Another related question is that of designing the right kinds of programming APIs or metaphors for interacting with different types of data -- this fundamentally depends on the conceptual model being used. Another big question that has emerged in recent years has to do with "trade-offs" among different resources.

Today's big data landscape features a range of complex computational environments, including many-core systems with multi-level cache hierarchies to distributed clusters with tens of thousands (or more) machines. As a result there are many different resources that can be traded off against each other, examples being memory usage, network communication, CPU instructions, cache misses, and so on. Most of these are new resources (in comparison with the traditional resources of space and time) in terms of our theoretical understanding of tradeoffs. Another dimension that has arisen is the acceptability of inaccurate or approximate answers. This leads to questions such as how many samples to use, how much time should be spent in verification, how much error is acceptable, and how these interact with the physical resource constraints. Finally, interpretability has emerged as an important theme in recent years. This includes issues such as framing questions appropriately, and understanding what it means to have an answer. Strategies for solving a specific problem are often different depending on whether the work is hypothesis-driven (question is posed first, followed by looking for evidence and designing algorithms for answering the question) vs exploratory (where we start with the data and try to design hypotheses).

Because of the foundational nature of this topic, theory typically precedes practice, often by many decades. Many theoretical advances from the past have had a major impact on the big data practice in recent years. Some examples include sublinear computations

on graphs; sampling and approximations theory; higher-order representations like graphs, tensors, and hierarchies; algorithms for multi-memory hierarchies; random projections; submodularity; non-convex optimization; and many others.

5.5.2 Major Challenges and Obstacles

This section discusses the different challenges and obstacles associated with foundational big data algorithms and theory. We group these challenges into three areas: abstraction challenges, data access challenges, and communication obstacles.

Abstraction Challenges: The major challenge in this area continues to be the tension between day-to-day work and abstractions -- in developing the foundations and the theory, one must work at higher-level abstractions, but for solving a specific problem in a specific domain, those abstractions may be less useful and specialized solutions might be more desirable. Another perspective is that we need to find the *right* abstractions that can capture the observed behavior of algorithms in specific domains.

Data Access Challenges: Navigating data access can be quite tricky -- without the right kind of data, it is difficult to understand what are the most effective and crucial foundational questions to focus on. A lot of researchers focus on what is called "Internet" problems, i.e., problems faced by big Internet companies like Google and Facebook. While those problems are important, more work needs to be done on hard questions in physical sciences, where it is also easier to get access to data. However, the challenge here is that interpretation of the data and results requires specific domain expertise and access to collaborations.

Communication Obstacle: Finally, communication between the researchers working on foundations and researchers working in application domains continues to be an issue, and raises significant barriers in developing new theories that may be applicable to emerging application domains.

5.5.3 Strategic Priorities

Some of the major open questions in this area also constitute the main strategic research priorities.

1. Although there is much work on understanding tradeoffs between different resources, research needs to be supported to improve our understanding and propel the development of better techniques that can efficiently navigate a large number of competing resource considerations. For example, when doing inference on large

amounts of high-dimensional data, how can we balance the desire for a few samples, low error, and efficient computation?

2. Research projects need to be supported that focus on developing algorithms, especially machine learning algorithms, for dealing with a large number of dimensions (features) and a small number of samples (e.g., in drug testing, or cancer research). How can one design algorithms and statistical methods in such a scenario and reason about the accuracy of those? A related challenge is that of doing inference in presence of a large number of labels, e.g., in personalization. Validation is also nontrivial when there is limited access to fresh data. We need methodologies to think about this question properly, and for doing evaluation when constrained by the amount of annotated data available.

3. Initiatives focused on understanding the consequences of complex human-machine interaction systems need to be supported across agencies. In systems where there are a large number of interacting agents, some of which may be computers and some human, there may be unintended consequences (e.g., discriminatory automated policies).

4. An important direction of research is the development of practicable theory of "causality". Although there has been much work on this topic, the practical impact of that work has been low. Funding both theoretical and development work in this area, as a partnership is an important direction.

5.6 Large-scale Inference & Learning

5.6.1 Current Landscape

There has been a tremendous interest in extracting insight from large-scale data sets collected from, for example, users' input, sensor readings, and scientific experiments. To perform high-performance inference and learning from these data sets is a nontrivial task, especially with constraints on computing and storage resources exist. To address this, the community has been exploring and developing novel methodological approaches to parallel and distributed computation for learning and inference. At the core, there has been significant progress in the fields of databases, high-performance computing, and hardware towards supporting large-scale stochastic optimizations. Notable examples include large-scale logistic regression and deep learning, which has benefited greatly from the development of new architectures, algorithms, and protocols that leverage the computing power of specialized hardware such as GPU and FPGAs.

These advances in the science of large-scale inferences and learning have led to successes in a wide range of application domains: computer vision, speech recognition, and natural language processing have been greatly improved by adopting deep learning; cancer genomics has made modest progress at a scale that it hasn't been able to do before.

5.6.2 Major Challenges and Obstacles

This section discusses some of the many challenges and obstacles associated with large-scale inference and learning. We group these challenges into three areas: data availability, bias and noise, and domain specific vs. domain agnostic challenges.

Data Availability Challenge: Academic data sets vary in their sizes and quality, but, as of May 2016, there is a lack of large enough data sets or benchmarks that are widely adopted/accepted by the community to evaluate the effectiveness and performance of proposed algorithms or systems. Those data sets would be easier to get from industry; however, industry either does not have the incentive or cannot release their datasets because of privacy concerns. It would be tremendously helpful to facilitate a data-sharing platforms and better partnerships between industry and academia.

Bias and Noise Challenges. Even if one has access to a large-scale data set, often the data are still largely unlabeled, which creates challenges for designed inference or learning algorithms to understand the bias and noise in the data set. For example, the

data collection process may reflect only a skewed perspective of the data, loss of accuracy may be introduced due to flawed data collection, etc.

Domain-specific vs. Domain-agnostic Challenges. Traditional learning algorithms such as SVMs are not necessarily optimized for certain domain-specific problems (e.g., in large-scale scientific computation), which leads to low performance or accuracy. Computer scientists have come up with better domain-specific solutions; we have seen many high-profile successful applications in the fields of speech recognition, computer visions, and computational genomics. However, there is still a gap between the advances in computer science and the awareness of these advances in a domain-specific context. An urgent effort is needed to help people of all disciplines, who are non-experts in computer science, know what is available to them.

5.6.3 Strategic Priorities

The breakout sessions considered the following areas as strategic priorities for advancing the innovation in large-scale inference and learning, as well as broadening understanding of the field.

1. Though there have been many high-profile successes in developing platforms dedicated for specific inference and learning tasks, it is still important to give priorities to combining “systems” (such as databases, HPC, hardware) and ML algorithms: fundamental progress in systems research may bring the inference and learning algorithms to a new level. For example, neural networks were proposed over two decades ago, but didn’t get much traction not only because of the unavailability of large-scale data, but also the lack of computing platforms that can support its demand of computing resources.

2. Machine learning has not evolved into a state where machine can automatically decide “the best” algorithm for specific inference and learning tasks on a specific dataset; currently, researchers need to experiment and choose from different algorithms to find the best algorithm and parameters. It is important to invest in research that advances the human-machine interactions for inference/learning, such as in mitigating the impact of bias and noise in data set.

3. Data driven inference and learning is starting to make decisions that humans cannot understand; for example, AlphaGo makes excellent moves that even human champions find hard to understand. Thus, another strategic priority is to make machine learning interpretable, e.g., in terms of knowledge of the processes generating the data, such that human not only receives the final result of inference

and learning, but also the intelligence and strategies that lead the learning to the result.

4. The training of non-experts in computer science is not prevalent in the graduate programs of different disciplines. For example, computational skills are not taught in a graduate class in biology; people have to self-teach themselves. Researchers outside of CS have difficulties in accessing and understanding these algorithms because they are written in a way that assumes significant background in computer science. It is import to develop classes to get non-CS researchers and students to understand what CS people are capable of making for them.

5.7 Learning Analytics and Education

5.7.1 Current Landscape

There has been an intense amount of work in recent years on improving learning and education using data-driven methods, and this remains one of the most important application areas for big data. Learning analytics is instrumental in supporting innovations in education, including individual and collaborative learning environments, MOOCs, intelligent tutors, learning management systems, open-ended learning environments, adaptive hypermedia, flipped classrooms, learning at scale, etc. It is also a key to personalized education, and through predictive modeling can help tailor education to the needs of a diverse student body. Much of the work in this area to date has focused on developing mechanisms to generate and collect assessment reports, e.g., through use of web-based intelligent tutors. Massively Open Online Courses (MOOCs) naturally generate large volumes of data, and there has been much work on analyzing that data, especially for studying dropout rates, student and instructor participation rates, etc. There have also been significant advances in building models of how students learn, and how scaffolding and adaptive feedback can influence their learning behaviors. Several visual analytics tools have also been developed for monitoring student progress over the duration of a course or even a degree.

5.7.2 Major Challenges and Obstacles

This section discusses the different challenges and obstacles associated with learning analytics and education. We group these challenges into three areas: the unclear potential of learning analytics, the lack of accepted models of the learning process, and the lack of data.

Unclear Potential of Learning Analytics: Learning is a complex issue and from a practical standpoint it involves multiple stakeholders starting from students with diverse backgrounds and their parents at one end of the spectrum, the teachers in between, and the school administrators and school boards at the other. On paper, they all have the same overall goals, but very different objectives and ways for evaluating the success of their goals. This means that they also have very different needs for analytics and mining methods, and the scales at which they need these analytics to be implemented. These stakeholders also currently lack the expertise to leverage the potential of small and big data and associated analytical techniques, especially since administering the data collection process, and understanding and acting on the analytics are not part of their training. Consequently, much of the data that has the potential for improving learning and education is neither properly collected nor analyzed in any meaningful manner. Stakeholder also lack knowledge of how to access analytics

to support or improve their learning since workflows and processes that can tie back directly to improving learning or making institutional change are missing. Finally, collection of large amounts of personalized data is not without risk, unless proper anonymization, access, and security measures are implemented at the school, district, state, and national levels.

No Accepted Models of the Learning Process or Educational Support: Learning is a highly interdisciplinary process involving insight from domain experts, education researchers, instructional design experts, cognitive scientists, neuroscientists, and learning scientists. Given the ubiquity of computers, and the potential of computational processes and the Internet, learning technologies and computer-based learning environments have also become essential tools in supporting learning processes. Given the diverse backgrounds of experts, it is still unclear how and what forms of learning analytics can make the most impact. On the one hand, analytics, due to their affordances to leverage multiple kinds of data, have the potential to support multiple ways of providing instruction and assessment, but a lack of readily acceptable frameworks, creates barriers in targeting different aspects and processes in learning, e.g., the method of acquisition of knowledge, the learning style (individualized, project- and problem-based or collaborative), and assessment to name a few. Clearly, there Additionally, it is also known that any given learning process is impacted by a host of factors at different levels – micro, meso, and macro – and the interactions across these levels are difficult to understand and hence to target for improvement. Researchers differ on fundamental questions such as how to define learning and how to define measures of success. Much of the analytics to date has focused on micro-interactions between instructors and students, or comparatively small-scale studies of individual and collaborative learning, and there is a lack of macro, cross-level, and cross-disciplinary understanding. More data driven methods (big data-based analytics) for hypothesis generation and verification, analysis and problem solving, can play a very important role in unraveling the mysteries and complexities of how people learn, and how to build better micro- and macro-level learning environments that span life-long learning.

Lack of Data Sets: Research in learning and education at all levels have been hampered by lack of access to rich datasets, which have been hard to assemble due to problems in employing state of the art technologies to the data collection and curation processes, as well as the private and protected nature of educational data. There are legitimate privacy concerns voiced by different stakeholders, but there is also additional barriers as many privacy regulations are not well understood or properly implemented. However, solutions may be imported from other domains, such as healthcare, to make meaningful advances in these areas. Other problems that impede advances, include: (1) variations in how data are collected across different institutions, making it hard to

aggregate the data, (2) lack of advancement in data collection methods for keeping up with current research that uses multi-modal data (e.g., verbal protocols, log files, facial expressions, and physiological data) to study and build integrated models of learning performance, learning behaviors, and self-regulation, and (3) no standardized formats for collecting and representing this type of data.

5.7.3 Strategic Priorities

While many different directions were discussed during the break out session, we focus on some urgent ones that we believe will have large impacts in advancing data mining and learning analytics to support advances in education:

1. Helping promote partnerships among researchers, educators, agencies, and other stakeholders to develop standards and a shared infrastructure for collecting multi-modal data and sharing data sets, analysis tools, and tool chains that support end-to-end analysis, including workflows that can be deployed in K-16 environments, and support researchers and practitioners. This would include developing and managing educational repositories, with privacy and security considerations, perhaps by federal and state agencies that facilitate rich data collection and data sharing to promote and facilitate cutting-edge research.

2. In parallel, inter-disciplinary stakeholders and researchers need mechanisms to scale up local efforts at specific institutions into global efforts across institutions.

3. We need to develop strategies to get domain experts and end-users involved in developing and deploying infrastructure with supporting tools, and continually improve this infrastructure as technology advances and research produces new results. We also need to raise awareness between domain experts, researchers, and end-users, so that people with different backgrounds can come together on these issues. One way to get communication started is to have a series of workshops discussing these issues at different education levels.

4. We need to advance the training of domain experts, practitioners, and other stakeholders in the education and learning arenas, so that they gain a better understanding of the potential of big data, mining, and analytics, work to deploy them in real settings, learn how to critically analyze and deploy research results and findings in ways that benefit the stakeholders, and improve learning and education at all levels.

6 Cross Cutting Recommendations and Priorities

The strategic priorities outlined in Section 5 focus on different areas within Big Data research. Here we consider the field as a whole and provide a set of recommendations and priorities that are high impact areas, areas of neglect, low hanging fruits, and important grand challenges. These recommendations come from discussions with workshop participants and have been further developed by the workshop organizers.

High Impact Priorities

While all the strategic priorities mentioned will lead to important advances with the field, there are a few that will have significantly high impact in terms of advancing research, innovation, and entrepreneurship. The four we focus on are the following:

1. Big data privacy and ethics guidelines: Without guidelines that researchers, corporations, and government agencies adhere to about data privacy and ethical uses of data, big data research will advance inconsistently. The guidelines are necessary to serve as a foundation for university regulations regarding the ethical use of big data. They can also serve as foundation documents for policy that can curb unethical uses of data by corporations and other large entities.

2. Data science lifecycle tools: For much of the population, data science is inaccessible. In order for it to grow more rapidly, we need to develop applications and tools that support the entire end-to-end data science pipelines: data collection, storage, cleaning, integration, machine learning, algorithms, and visualization. These tools need to be integrated into usable, accessible suites for data analysis. Doing so will increase data driven decision making and let more people take advantage of the advancements within the field.

3. Funding mechanisms for supporting early stage interdisciplinary partnerships & transition to larger grants: One very clear message from the workshop is that interdisciplinary partnerships are hard and take time to cultivate. Because of the silo structure of academic institutions, agencies like NSF need to provide funding mechanisms for bringing groups together to identify synergies and write both small innovation grants and large-scale ones. There need to be small grants to help begin the collaborations and then larger follow-up grants that can be applied for once the ground work has been laid.

4. Algorithm explainability: Automated data-driven decision-making tools are being increasingly used in all application domains; however, in many cases, the decisions made by those tools cannot be easily explained. Similarly, we often have complex machine learning models that are effective at a prediction task,

but are very difficult to interpret and understand. More resources need to be devoted to synergistically develop theories and systems that tackle these fundamental problems.

5. Robust learning and decision-making: We need better understanding of how robust our learning and analysis algorithms are, and under what scenarios they work well. In particular, more research needs to be devoted to the scenarios where the amount of labeled/annotated data is small relative to the number of dimensions or parameters, where there is significant skew or bias (potentially unknown) in the training data, and where the number of different labels/classes is very large.

Neglected Areas of Research

When a field moves this rapidly, there are always areas that get more attention than others. Some areas of neglect that were mentioned at the workshop included the following:

- Data transparency
- Data reliability and trust
- Data provenance
- Large-scale systems building
- Visualization and visual analytics

Systems and visualization used to have dedicated programs that are no longer in CISE. Data transparency, reliability and trust are all new areas that have not seen as rapid research advancement. Data provenance has been an active area of research in some domains, but lags in others. Putting resources in these areas is important to the big data field as a whole.

Low Hanging Fruit

A few simple mechanisms that would help researchers in this area are the following:

- Researchers at the meeting were excited to learn about big data hubs. For many of them, it was their first exposure to them. Data and infrastructure sharing can be difficult. However, if big data hubs can be given the resources to facilitate this, it would help further propel big data research.
- Giving interdisciplinary research projects longer grant cycles will help them learn each other's domains and make more research progress. Four to five year grants for those projects seem more appropriate than three years.
- The majority of universities cannot afford to have multiple large-scale infrastructures for research and development. While some cloud computing

options are inexpensive to academics, the range of different infrastructures is not readily available. Setting up a mechanism for sharing or crowd sourcing infrastructure as well as other research resources will enable small business and smaller universities to engage in big data research and development.

Grand Challenges

Large inter-agency grand challenges are another way to bring together experts from different domains. Below we list some possible grand challenge ideas and agencies that may work with NSF to support projects in these areas.

1. Protocols for secure, private, and ethical sharing of data (NITRD AGENCIES)
2. Self-correcting algorithms that identify and deal with bias (DHS, DOD)
3. Public knowledge-bases that can be used to advance artificial intelligence applications (NITRD AGENCIES)
4. Using big data to address societal scale problems:
    - Making communities more resilient (DOE)
    - Helping understand movement of refugees (State)
    - Reducing human trafficking (State)
    - Improving transportation (DOT)
    - Stability of financial markets (OFR)
    - Impact of automation on jobs (DOL)
    - Using technology and learning analytics to advance education in poor regions of the world (DOE, State)
    - Climate change (NASA, DOE)
5. Using big data to address health and science challenges:
    - Advancing personalized medicine (NIH)
    - Improving global health (DHHS, State)
6. Developing a rapid “scientific innovation cycle” that incorporates big data to inform policy (DHS, DOD, DoC)
7. Anonymizing data of different age student groups (elementary, middle school, high school, college) and setting up teams of educators and interdisciplinary researchers to develop theory and analytics for improving learning (DOE)

7 Lessons Learned

Hindsight is always 20/20. Here are some recommendations for the next team that runs a PI meeting.

1. To ensure that turnout is high and hotel options are larger, meeting planning should begin one year in advance.
2. Because the group present spans a large number of fields, it is important to have enough time for ice breakers.
3. Some of the keynote talks should be given by domain experts, that is, not all by computer scientists.
4. Many participants book their flights early in the afternoon on the second day. It may make sense to have a workshop that is only 1 ½ days.
5. Having options for online questions and comments may have been useful for discussions.
6. Write out a draft of the final report at the end of the meeting so that it will get done faster and there will be more active participation.

A G E N D A

WEDNESDAY, APRIL 20, 2016

8:15 AM – 9:00 AM Registration; continental breakfast

9:00 AM - 9:30 AM Welcome & Workshop Goals

Lisa Singh, Georgetown University
Jim Kurose, National Science Foundation (CISE)
Chaitan Baru, National Science Foundation (CISE)

9:30 AM - 10:30 AM Big Data Regional Innovation Hubs Panel

Northeast Hub
René Bastón, Columbia University

South Hub
Renata Rawlings-Goss, Georgia Institute of Technology
Lea Shanley, University of North Carolina at Chapel Hill

Midwest Hub
Melissa Cragin, University of Illinios at Urbana-Champaign

West Hub
Meredith Lee, University of California, Berkeley

10:30 AM - 10:45 AM Break

10:45 AM - 12:00 PM Emerging Big Data Topics

Vipin Kumar, University of Minnesota
Big Data in Climate: Opportunities and Challenges for Machine Learning and Data Mining

danah boyd, Microsoft Research
Interpretation Pitfalls in Big Data

Anthony Townsend, Bits and Atoms
Cities of Data: Making Sense of the New Urban Science

12:00 PM - 1:30 PM Lunch (provided by workshop)

Jason Schultz, Office of Science and Technology Policy

WEDNESDAY, APRIL 20, 2016 (CONT.)

1:30 PM - 3:15 PM Breakout Sessions - Part 1

3:15 PM - 3:30 PM Break

3:30 PM - 5:00 PM NSF Funders Panel

Reed Beaman, Biological Sciences (BIO/DBI)
John Cherniavsky, Education and Human Resources (EHR/DRL)
Nandini Kannan, Mathematical Sciences (DMS/MPS)
Sylvia Spengler, Computer and Information Science & Engineering (CISE/IIS)
Chengshan Xiao, Engineering (ENG/ECCS)
Heng Xu, Social, Behavioral & Economic Sciences (SBE/SES)
Eva Zanzerkia, Geosciences (GEO/EAR)

5:00 PM - 5:15 PM Poster and Demo Setup

5:15 PM - 6:30 PM Poster and Demo Session

6:30 PM Evening Reception and Dinner

Table Talk

THURSDAY, APRIL 21, 2016

8:15 AM – 9:00 AM Registration; continental breakfast

9:00 AM - 9:15 AM Opening Remarks

Chaitan Baru, National Science Foundation (CISE)

9:15 AM - 10:25 AM Asian International Partnerships

Etsuya Shibayama, University of Tokyo [slides]
Takeaki Uno, National Institutes of Informatics [slides]
Satoshi Matsuoka, Tokyo Institute of Technology [slides]
Hayato Yamana, Waseda University [slides]

10:25 AM - 11:20 AM Breakout Sessions - Part 2

11:20 AM - 11:30 AM Break

11:30 AM - 1:30 PM Breakout Session Presentation & Working Lunch

THURSDAY, APRIL 21, 2016 (CONT.)

1:30 PM - 3:00 PM Federal Agency Data & Funders Panel

Chaitan Baru, National Science Foundation (NSF) [slides]
Allen Dearry, National Institutes of Health (NIH) [slides]
Stephen Dennis, Department of Homeland Security (DHS) [slides]
Daniel Duffy, National Aeronautics and Space Administration (NASA) [slides]
Charles Fay, Department of Transportation (DOT) [slides]
Greg Feldberg, Department of Treasury (OFR)
James St. Pierre, National Institute of Standards and Technology (NIST)

3:00 PM - 4:00 PM Plenary Discussion

4:00 PM - 4:30 PM Concluding Remarks