Challenges and Opportunities in Big Data Research: Outcomes from the Second Annual Joint PI Meeting of the NSF BIGDATA Research Program and the NSF Big Data Regional Innovation Hubs and Spokes Programs 2018

Samarth Swarup, Vladimir Braverman, Raman Arora, Doina Caragea,
Melissa Cragin, Jennifer Dy, Vasant Honavar, Heng Huang,
Ryan Locicero, Lisa Singh, Chris Yang

September 10, 2019

Introduction

NSF is funding research through its programs in Big Data to cover all aspects of the Big Data ecosystem. These research programs cover mathematical, statistical, and algorithmic foundations, the development of new technologies and techniques, hardware and software infrastructure-building, and applications requiring innovative uses of Big Data.

The second annual meeting of the principal investigators (PIs) of the NSF BIGDATA Research Program, the NSF Big Data Regional Innovation Hubs Program (BD Hubs), and the NSF Big Data Regional Innovation Spokes Program (BD Spokes) was held at the Westin Alexandria Hotel in Alexandria, VA, on June 20-22, 2018.

The purpose of the BIGDATA research program is to, "develop new methods for deriving knowledge from data; construct new infrastructure to manage, curate, and serve data to communities; forge new approaches for associated education and training."

The Big Data Hubs and Spokes were created by the Computer and Information Science and Engineering (CISE) directorate as part of the administration's National Big Data Research and Development Initiative. The goal of the initiative is to promote research and development in methodologies and infrastructure to extract information from the large and complex data sets that are being routinely generated in many different domains today.

The annual meeting of all the PIs serves to facilitate the transdisciplinarity envisioned in these programs, to increase awareness of the breadth of research and infrastructure development in these programs, and to promote collaborations. There were more than 200 attendees over the course of the three days. Approximately 160 people attended the BIGDATA Research sessions on June 20, approximately 180 people attended the joint sessions on June 21, and approximately 84 people attended the Hubs and Spokes sessions on June 22. There were 6 Government agencies represented, ∼100 Universities represented, and over 130 projects shared by PIs in lightning talks and poster sessions.

The goals of the meeting were to promote interaction between the participants in the programs, to facilitate collaborations, to increase awareness of related funding opportunities, to facilitate interactions with industry partners, to provide exposure to students to the latest research in Big Data, and to help situate the two programs within the larger Big Data ecosystem.

Table 1: Session Titles

Foundational Big Data Algorithms
Big Data and Health/Medicine
Big Data, Learning Analytics, and Theory
Large-scale Inference and Learning
Using Big Data for Social Good
Big Data and Smart Cities
Forming Interdisciplinary Partnerships
Birds of a feather: The Cloud
Birds of a feather: Data-sharing

The program for the meeting included seven breakout sessions and two birds-of-a-feather sessions. The topics for the sessions are listed in Table 1. In each session, participants discussed research directions, challenges, and opportunities relevant to the topic of the session. A number of common themes emerged from these discussions. We summarize these below.

Insights and Recommendations

The breakout sessions were organized around four common questions:

- What are the most important challenges for the next 3-5 years?
- What are the most impactful opportunities in the next 3-5 years?
- What would you need in order to take advantage of the opportunities and address the challenges?
- How can the Big Data Hubs, NSF, or the people in the room contribute to the other questions?

The birds-of-a-feather sessions were not organized around specific questions, but encouraged a discussion around common interests and the topic of the session. In the following, we have synthesized the comments and responses to each of the above questions.

Challenges:

While the traditional Big Data concerns of volume, velocity, variety, and veracity were also raised, much of the discussion focused on a number of new challenges. These challenges span the full range of the data processing pipeline, from data collection to societal impact. We have organized these into stages and discuss them in order below.

Data Collection:

- *Data availability*: In many cases, there is a lack of direct access to ground truth data, which affects the quality of the output. Usually the data we have access to contain biases or noise. For example, many data sources, such as health and medical data, still have to deal with low quality labels and many data even lack necessary annotations, making it difficult for data scientists to use these big data. On the other hand, many data scientists need to work closely with biomedical experts to know the real needs of the health and medical domains.
- *Data quality*: Sometimes the input we receive are adversarially generated. Even when that's not the case, they may not be neutral as the persons collecting the data might be collecting only the part of the data which is more important for their study. Such things happen a lot in many field, such as social sciences. So, in order to better use the data, it is important to understand the incentive of the agent who collected the data.
- *Data are more observational than experimental*: Observational data, in addition to problems of unknown bias and noise, have unknown causal processes that generate them, as opposed to experimental data generated in carefully controlled circumstances. While the latter may be preferable in many cases, one of the main challenges here is to how to make the experiments feasible in terms of computation and resources. Alternatively, we are faced with the challenge of developing meaningful causal models for observational data.
- *Lack of theory of the data collection process*: Many data sets that are collected and used reflect a snapshot of the time when the data was collected, and may not be very relevant for a longer process. Further, it is impossible to get a complete snapshot of data for a large open system such as a city.
- *It is not well understood what is a good point when one should change their knowledge base*: As the world evolves, the knowledge we have extracted from the data becomes stale and requires updating. Determining when and how to do this is a challenge. We need to bring people from domain sciences and data science together to better understand when a data set becomes stale and is not useful to make a predictive model.

Data Organization and Control:

- *Privacy and security*: Data privacy and security issues have been discussed for many years, but they are still one of the most important challenges for most domain sciences, such as biomedical science, social science, and more. We need better algorithms for de-identifying the genomic sequencing, electronic medical records, and many other biomedical data for privacy-preserving data publishing and secure data mining. The existing differential privacy algorithms don't work well for the biomedical data so far.

- *Ownership and access*: Ensuring rights of the entities from whom the data are generated is a challenge, as it is often the case that the entities controlling access to the data are different. Restricted access also raises reproducibility and verifiability challenges.

- *Human values*: In addition to the above, a continuing challenge is how to design systems that respect a range of fundamental human values, such as transparency, fairness, dignity, safety, and more.

Data Processing Methodology:

- *Hardware*: Algorithms are designed to work for any system, making them sometimes slow on particular hardware. Co-designing hardware and algorithm so that the algorithms are more customized to a particular hardware is a challenge. For example, one practical obstacle is the synchronization between GPU and computer. Hardware and algorithm design can be done together to improve scalability. For example, in the field of large-scale learning and inference, there has been a lot of attention in the recent past to design distributed learning algorithms, but the field is far from being mature.

- *Combining top-down and bottom-up approaches*: Theory generally takes a top-down approach by looking for particular kinds of structure in the data (which may not necessarily be present): we first design algorithms and then test it on data. On the other hand, many fields such as data mining take a bottom up approach—collect data and then try to infer the useful aspects of the data. Combining the two viewpoints to generate meaningful insights is a challenge.

- *Data integration/fusion*: In many domains, the kinds of data being generated are quite heterogeneous. For example, in the health and medical data, we have genome sequences, pathological images, electronic medical records, body sensing data, and more. It is challenging to integrate them for data science research. We especially need to integrate domain knowledge of different types of data in doing data fusion. This knowledge is contextual, i.e., not always present in the data themselves. A major challenge is how to identify this kind of knowledge and use it in data integration.

- *Extracting structured meaning from unstructured data*: Text mining is important for many domains. For example, in biomedical data science research, various kinds of textual data are being generated, such as unstructured electronic health records, PubMed and other web data, lab reports, clinical notes, etc. Designing domain-focused text mining tools and natural language processing techniques to effectively analyze these big biomedical text data is a challenge.

The Big Data Ecology:

- *Cross-disciplinary collaborations*: An important challenge is to include include domain experts into data science and use their expertise to create the interpretable results.

- *Recognition*: Initiatives around social good, public good, data journalism, value-sensitive design, FAT*, and more are a very important part of the Big Data ecology, Many academic departments are still silos, and research in the above fields is very interdisciplinary – it does not fit into any traditional computer science or statistics research area. There is a concern within many departments about the quality of the research in these areas, and more specifically, the depth of the innovation for a particular discipline. This siloed view of research is a challenge for these areas and can have a direct impact on promotion and tenure of junior faculty. Also, to work on these issues "part-time" can be difficult given the effort needed to engage in this type of work. This in turn means that many researchers avoid working on these topics because of the time commitment given other demands on their schedule.

- *Community cohesion*: A number of different initiatives related to data science for the above topics have been underway for a few years. However, there is still a lack of cohesion among various efforts, which leads to redundancy and inefficiency. Everyone is still "reinventing the wheel". While some of these groups have connected on their own, e.g., University of Chicago and University of Washington, many have grown more independently, e.g., Bloomberg. To avoid reinventing everything from scratch, templates and lessons learned need to be aggregated and shared. Foundations, government agencies, and those in the public sector are unaware of different programs that already exist, as well as the gaps that could use more research, development, and growth. In general, there is also a large interest in this from many groups, including students and project providers. We need to learn how to leverage all the excitement.

- *Funding support*: While there are vehicles for a small number of large interdisciplinary projects, there are not as many vehicles for smaller projects. Many of these projects undergo the same review process, without recognition for the particular needs and constraints of emerging cross-disciplinary fields.

Societal Impact

- *Lack of funding for research in social good and related fields*: Researchers always need more funding. However, in this domain, funding is more complicated. These types of projects have both a research and a practical component, and very often, the research is less impactful than the practical component. Industry is not heavily involved because social good in any form can be difficult to monetize and other incentives may not be obvious for them to get involved. In terms of government funding, if PIs can get initial funding for the proof of concept, it is difficult for them to get continued funding for the final implementation. Less support exists for implementation in real world settings. Going from research to implementation/practice currently requires reapplying for funds. Finally, impact studies also need funding. Understanding longterm impacts require longitudinal studies that are typically costly to employ.

- *There are not enough translators between researchers and those in the public good space*: People who work in research do not always understand the public sector world and people who work in the public sector do not always understand the research world. It can take significant time for everyone to learn to talk to each other and translate. There are practical issues that researchers may not understand or have exposure to. There are also different types of people who are important at different stages through the life cycle of these projects. For smaller projects, many of those roles are being handled by those who are not experts in them.

- *It is hard and takes time to build trust with vulnerable populations/communities*: It takes time to build a research community around a particular social issue. But researchers who are not connected to the issue directly may not understand this commitment or have the support to undergo a longterm relationship with those that need help. Many times, the domain experts work directly with people who are being helped, but others on the research team do not have direct contact with the people they want to help. This can reduce their level of engagement and commitment to the project.

- *Technology seems to be increasing divisions, rather than reducing them*: There is a growing perception among the general public that technology is increasing social polarization, socio-economic divisions, and health (and other) disparities. One of the biggest challenges facing us today is how to orient technology towards reducing divisions, and how to educate the general public in participating in the beneficial use of technology.

Opportunities

Participants discussed a number of opportunities for enabling innovative Big Data research in the near future. These can broadly be organized into new kinds of data sources, collaborations between academia and industry, new training programs, and new research infrastructure.

New kinds of data sources:

- *Medical imaging data*: A vast quantity of high-resolution medical imaging data are now being generated every year. For example, fewer than 500 digital pathological images have more image pixels, i.e., image content, than all images together from the ImageNet database (more than 14 million regular pictures). Such massive health data impose an extremely difficult computational challenge and also a new opportunity for medical analysis and understanding that did not exist before..

- *Widespread adoption of electronic health records*: Due to the widespread adoption of electronic health records, there are great opportunities for researchers to diagnose and treat patients better and improve health quality.

- *Non-traditional sensing techniques*: Novel sensing techniques, such as wearables, provide a unique opportunity for low-cost everyday healthcare for patients and aging people.

- There are several impactful research topics, such as precision health and epidemiology, which greatly need the help from Big Data research.

Collaborations between academia and industry:

Increased collaboration between academia and industry will be mutually beneficial and offer opportunities for new kinds of fundamental and applied research at large scales. Representatives from Microsoft, IBM, Amazon, and Google attended the meeting and described multiple initiatives, programs, and incentives for collaboration offered by them.

- Microsoft announced the launch of the Microsoft Data Repository, which hosts data that all Microsoft researchers have produced.

- Microsoft is investing into Big Data Cloud Collaborations with the NSF Big Data Hubs

- IBM Watson: Any user can build a platform through their developer account

- IBM is launching a research community around quantum computing (IBM Q, QISKit on GitHub)

- IBM also has a Blockchain developer center.

- IBM academic initiatives provide access to Cloud, Courses, and Software via `onthehub.com/ibm/`.

- Amazon Web Serives (AWS) is making resources available through partnerships with NSF and otherwise, such as Cloud credits and the Neptune graph database service.

- Google is also giving Cloud credits through their GCP research credits program. Faculty researchers in eligible countries can apply.

New training programs

- Participants noted the emergence of *interdisciplinary training programs* to create the next generation of domain-focused data scientists.

- Participants also noted the increased focus on *workforce development* in various academic departments and schools.

New research infrastructure

- *Regional cyberinfrastructure*: New regional cyberinfrastructure is slowly beginning to emerge. Expanding the scale of and access to these systems and services is an opportunity to stimulate Big Data research.

- *Development of "cooperatives", such as EarthCube and DataCenterHub.org*: These and other similar organizations focused on data preservation, sharing, and access are creating new opportunities for large-scale data-centric and problem-centric science.

What is Needed:

The participants identified a broad set of “needs” for enabling progress in Big Data Science in the near future. Some of these are institutional, in the sense that they are desirable changes at the level of departments, universities, government agencies, and funding agencies. The rest we have grouped under the heading “general”, because they refer to larger social and cultural matters.

Institutional

- *Joint education between CS and health/medical programs*: Joint courses, certificate programs, and joint degrees were suggested as essential for developing the next generation of data scientists. This applies not just to health/medical programs, but other domains as well. The “need” is for data scientists who are equally deeply trained in the domain science as they are in data science.

- *MOUs with cities can facilitate Big Data Urban Science resaerch*: Urban Science was pointed out as one of the paradigmatic examples of Big Data research applications. There is a growing need for partnerships between academic institutions and cities to facilitate this research. Memoranda of Understanding (MOUs) are a mechanism to help achieve this kind of collaboration.

- *Incentives for faculty to engage*: There is a need to new kinds of incentives and structures to encourage faculty to engage with the community for applied Big Data research. For example, more universities should have Extension Offices instead of having teaching faculty do extension.

- *Sustainable solutions*: The path from research to application can be quite long. It is necessary to have the means to maintain continuity of relationships and institutional knowledge during this process.

- *Models for scaling up*: We need better models for scaling up the results from academic research and proof-of-concept demonstrations to real-world applications.

- *Improve research relationships across campuses and states.*

- *Improving the review process*: The review process should have mechanisms for recognizing the impact of building best practices for different types of interdisciplinary groups, and the impact of trying to connect state of the art ideas across fields and into practice.

- *Guidance material for working with cities*: For many researchers in engineering fields, there is a lack of experience in working with cities, bureaucracies, and operational agencies. There is a need for training and guidance material for enabling Big Data researchers to engage in such collaborations.

- *Shared spaces*: One suggestion for enabling collaboration was to have researchers embedded in shared spaces with city government, e.g., exchange programs.

- *Standardized data sets*: There is a need for standardized and benchmark data sets to minimize the effort of gathering data for methods development, and to enable sound comparison of methods. Possibilities include national/federal data sets, “random social media sampling”, consultants data that may be “self-serving”, city-collected data, and more.

- There was a call for *increased regionally-accessible cyberinfrastructure* There was, for example, a sense that there is a need for more equitable access to resources that would result in more general capacity building and the opportunity to “raise all boats; this would not only improve (scientific) production, but encourage increased participation by researchers and students”. While this wasn’t an explicit statement, an idea that was repeated was that there would be more realized benefit from existing CI investments and resources if there was improved cross-institutional collaboration.

General

- We need *strong integration between data science and domain science.* One suggestion was to put together people who think differently and are from different domains and then give them some specific issues to discuss. Further, the participants noted that every NSF proposal requires a data management plan, and that perhaps, for Big Data proposals, we can also require a section that establishes a real connection between the proposed research and a real-world task in data science.

- *Learn from other domains that cross disciplines*: It was also recognized that several of these challenges are not new; they have been addressed before in different contexts and at different scales. Thus there is a need to learn from the experiences of other domains, e.g., Urban Planning.

- Developing and promoting *partnerships between scholarship and activism* is vital, especially for emerging fields of application such as social good, public good, data journalism, etc.

- *Communication*: Societal communication would benefit Big Data research in multiple areas, such as health and medical research. Improved communications can create positive environment for Big Data research.

- Opportunities to engender *new, cross-sector and multi-institution collaborations*, and to address the sociotechnical challenges that inhibit or slow the production of science and scholarship. Noted, also, was the importance of physical (in-person) meetings, for real-time, face-to-face discussion and brainstorming – often the launch pad of new collaborative relationships.

- *Public participation and hackathons*: Participants noted the importance of letting citizens use data sets and letting some of those data sets be used for hackathons. In addition to driving innovation, this would also help build a broader scientific and technological culture.

How NSF, the Big Data Hubs, and Others Can Contribute:

Participants identified and discussed a number of ways in which NSF, the Big Data Hubs, and other institutions can help give greater impetus to Big Data research.

- *Joint NSF-NIH programs*: Given the different kinds of focus NSF and NIH programs typically have, joint programs would help to create Big Data projects having both transformative and translational research values.

- *NSF Research Coordination Networks* program could connect related and complementary NSF and NIH programs.

- *EAGER and RAPID grants* will help the interdisciplinary collaborations to successfully create new joint Big Data research projects.

- NSF may take the lead in *facilitating collaborations* with NIH, stakeholder, government, and companies.

- *Increased links* with other types of organizations, such as professional societies, in order to increase interdisciplinary partnerships.

- *Build fellowship programs* with PhRMA and other government agencies.

- NSF needs to have *funding streams beyond the traditional ones* that focus on interdisciplinary team training, interdisciplinary content training, ethical frameworks training, and dissemination beyond papers into practice. This will help incentivize working outside of a traditional disciplinary silo. The Spoke program is one good example of this type of funding mechanism, but very few proposals are funded through that program. Also, the Spokes need to have a connection to the mission of the Hubs. That can limit the types of submissions. Promoting mechanisms that exist at NSF and can be used for social good type grants, e.g. Convergence, data literacy for all programs, is important.

- While there are vehicles for funding a small number of large interdisciplinary projects, there are not as many vehicles for smaller projects. Also, many of these projects undergo the same review process; however, because the outputs are different, this should not be the case. The review process should have mechanisms to not only reward the research impact, but the *impact of building best practices* for different types of interdisciplinary groups, and the *impact of trying to connect state of the art ideas across fields and into practice.* Finding ways to support and reward the effort and overhead associated with cross-disciplinary research and implementation related to larger societal scale issues is important. Perhaps there can be a required section about accomplishments in non-traditional research as a way to incentivize this type of work.

- NSF can *support more cross-disciplinary conferences* that are devoted to best practices, templates, measures, and training modules for successful interdisciplinary research. Examples are the Bloomberg Data for Good Exchange (`https://www.bloomberg.com/company/d4gx`), and the KDD Impact Program.

- Since data science for social good is attractive to students, help *spearhead and fund initiatives* that promote this in classrooms or externally, and connect these initiatives to existing programs and new programs.

- The Hubs can also engage in developing *working groups* that think about issues related to Big Data and Social Good and related efforts.

- NSF and Hubs can help *foster and sustain relationships between researchers and cities.*

- The current NSF BIGDATA Hubs and spokes programs are helpful to address some challenges, such as *linking different experts*, *forming workshops* to facilitate the interdisciplinary collaborations, etc.

- NSF and Hubs can help with *awareness-building of technology and data science techniques* (what it can do, ROI).

- There was a call for better ways to *increase cross Spoke (project) awareness*, and to *encourage potential cross-project collaborations*. Further, the BD Hubs could facilitate better access to shared tools and more "general solutions." Interestingly, it was noted that the increase in technology seems to be increasing "silo-ing" in academic research, rather than decreasing this; and, that the BD Hubs might be well positioned to address inertia in this silo-ing.

- A noted concern was the lack of "community standards," which seemed to generally mean data and sharing standards. This was another topic for which there was real interest for the BD Hubs to facilitate. Further, where standards exist, the Hubs could *make standards available for academic and industry.*

- *Practitioner training, executive training*: To move Big Data research results beyond academia, NSF could facilitate training for practitioners and executives who are the end-users of the new technologies being developed.

- *New funding formats*: Longer-term grants. For example, 5-10 year grants as needed for longevity, relationahip-building and sustaining (LTER model); 2-3 year grant cycle is damaging.

- *Create partnerships with other agencies*, such as EPA, NOAA, HHS, to address local issues and research challenges.

- In addition, the *need for increased cross-BD Hub communications* was identified, to increase awareness of the events, activities, and opportunities across the Hubs & Spokes program and related projects; this should include "mapping" of funding opportunities.

- *Map funding opportunities* across federal agencies and foundations, and increase awareness of these opportunities.

Conclusion

This was the second year in which the annual PI meetings for the BIGDATA program and the Big Data Hubs and Spokes programs were held together. Participants consistently reported that they enjoyed this expanded format with greater interaction leading to more opportunities for collaboration, and greater awareness of the breadth of research being supported by NSF. It was also important to have continued industry participation in the event and to have attendees from other government agencies. Several students also participated, and gained experience about the funding process and learned about the breadth and depth of Big Data research.

In addition to the authors, many other participants contributed to the discussions, panels, and talks. The event was a success due to all their efforts. We gratefully acknowledge all their contributions.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1834405.