NSF BIGDATA PI Meeting – Domain Specific Research Directions and Data Sets

Lisa Singh, Amol Deshpande, Wenchao Zhou,
Arindam Banerjee, Alex Bowers, Sorelle Friedler, H.V. Jagadish,
George Karypis, Zoran Obradovic, Anil Vullikanti, Wangda Zuo

Fall 2017

1 Summary

In 2017, PIs and co-PIs that were funded through the NSF BIGDATA program were brought together along with selected industry and government invitees to discuss current research, identify current challenges, discuss promising future directions, foster new collaborations, and share accomplishments, at BDPI2017. Given that two recent NITRD [1] and NSF [2] meeting reports contained a set of recommendations, grand challenges, and high impact priorities for Big Data, the organizers of this meeting shifted the focus of the breakout sessions to discuss problems and available data sets that exist in five application domains – policy, health, education, economy & finance, and environment & energy. These domains were selected based on a survey of the PIs and co-PIs during the meeting registration process. The survey contained different application areas highlighted in previous meetings. Before summarizing key domain-specific ideas identified by the breakout groups, we begin by highlighting common big data research concerns and challenges. More detailed slides that were presented by the different breakout group leaders are available at `https://www.bi.vt.edu/nsf-big-data/`. In general, this report serves as a blueprint for promising big data research in five application domains.

2 Common Big Data Research Concerns and Challenges

While there were a number of unique domain-specific challenges, there were some broader concerns that were repeatedly mentioned across the breakout groups. Here we focus on two of them because they have a large impact on advancing big data research more broadly.

Concern 1:

How can successful multidisciplinary collaborations be facilitated given the potential misalignment in training, perspectives, and research goals?

The ramp-up time for interdisciplinary research can be substantial. There are discipline-specific language barriers that need to be overcome. Researchers from different disciplines need to teach each other about their domain, methods, etc. They all also need to have an interest in advancing research in other disciplines – too often domain experts are viewed as clients and computer scientists and statisticians are viewed as programmers and data janitors. In reality, they are all researchers who need to view each other's disciplines as equal – otherwise the research partnership is doomed from the start.

Recommendations:

1. Allow awards to include time for interdisciplinary training of team members during year one. In other words, require new collaborations to develop training lectures that can be shared on project websites as an outcome of year one.
2. Require students across disciplines to be supported on interdisciplinary grants – to promote interdisciplinary thought and training of students. This will help alleviate some of the training mismatch for the next generation of researchers.
3. Develop workshops that can be disseminated through Hubs and Spokes for successful interdisciplinary collaborations.

Concern 2:

How can we increase the availability of high quality data sets?

For many of the challenging, societal-scale issues, clean, well-processed data do not exist. Collecting, transforming, labeling, and validating these data for further analyses is costly and time-consuming. However, these pre-processing steps are necessary to ensure high data quality and eventually, meaningful big data results.

Recommendations:

1. Support grants that focus on developing pre-processing tools that can be shared with the community and easily adapted for different domains.
2. Generate more universally accepted quality standards for big data so that researchers understand the limitations of the data without wasting time going through them.
3. Develop privacy-driven pre-processing tools that identify unique records or data features that should be altered or not shared to protect privacy.

4. Have calls focused on data sharing and support infrastructure costs associated with the data sharing. Create and maintain benchmark databases that are publicly available for research.

3 Domain-Specific Big Data Research Directions and Data Sets

In this section we present research directions and available data sets identified in the breakout sessions. The research directions focus on research that is promising in the short-term, i.e., a three-year window. The domains presented in this report are: policy, health, education, the economy & finance, and the environment & energy.

3.1 Education

There have been a number of recent successes in this domain, including degree planning software for identifying early warning of poor student performance, intelligent tutoring systems, and educational analytic reporting tools. Over the next three years, promising directions for advancing the state of the art include:

- Develop a sharing environment that can be used to combine and run models on restricted data that are siloed across the industry. This infrastructure should allow researchers to provide analytics, test different methods, and understand outcomes across the combined data sets. The infrastructure should be developed with student privacy as a central tenet of the initial design.

- Develop partnerships among workforce development specialists, data scientists, and education specialists to model career paths and provide recommendation software for students and adults at different stages in their career.

- Create data sets, analytics, visualizations, and learning algorithms that analyze the learning life cycle – from how teachers teach different topics to student engagement to student learning outcomes to career outcomes.

- Develop clear usage guidelines related to the confidentiality, ethical uses, and privacy of these data.

While the number of available data sets is limited, there are a few projects that have anonymized and released data for use by researchers. A few data sets exist that are related to student performance and assessment on different Massive Open Online Courses (MOOC) platforms – the 2010 KDD Cup Challenge data set that contained interaction records between students and a computer aided intelligent tutoring system for learning algebra,[1] the 2015

[1] KDD Cup Challenge (2010): https://pslcdatashop.web.cmu.edu/KDDCup/downloads.jsp

KDD Cup Challenge data set that contained data about student interactions with the virtual learning environment (China's XuetangX MOOC platform),[2] and the Open University Learning Analytics dataset (OULAD) that contains anonymized student demographic and interaction data for over 30,000 students across 22 courses.[3] These data sets can used to begin to understand student learning, student performance with tutoring guidance, general learning styles, knowledge requirements for progressive learning, etc.

There are also data sets that can be used to explore research questions related to adaptive learning. For example, Rossi and Gnawali released anonymized versions of discussion threads from 60 Coursera MOOCs.[4] Papousek and colleagues have released a data set related to student learning of geography facts.[5] Finally, there is also a substantial literature that is being developed about learning analytics methodologies. SoLAR curates a data set containing these different research sources to support computational analyses,[6] and the AFEL Data Catalog contains a collection of nonuser-centric data sets for understanding different online and social learning contexts.[7]

Together, these data sets provide initial testbeds for exploring different research questions related to online education.

3.2 Environment and Energy

Big data analytics has been successfully applied in the broad field of environment and energy. It has been used to study the air quality with a large amount of air quality sensors, water resource management (e.g., water supply, water quality and quantity), smart cities (e.g., smart transportation, smart parking, smart buildings), climate changes in terms of its courses and potential impacts, environmental impact on food, building energy efficiency standards and policy, smart grid enabled by the smart meters, ecology and ecosystem management, as well as geophysics (e.g., oil and gas exploration and production, geothermal, contaminant transport, carbon sequestration, and ground water).

Over the next three years, promising directions for advancing the state of the art include:

- Develop better, large-scale visual analytics methods and tools.
- Develop hybrid analytics approaches that use cloud analytics for large, public data sets and local analytics for sensitive data sets.
- Develop approaches and tools for interfacing machine learning with scientific models.

[2] KDD Cup Challenge (2015): http://data-mining.philippe-fournier-viger.com/the-kddcup-2015-dataset-download-link/

[3] OULAD: https://www.nature.com/articles/sdata2017171

[4] Coursera MOOC discussion threads: https://github.com/elleros/courseraforums

[5] Adaptive learning: https://github.com/adaptive-learning/data-public

[6] SoLAR: https://solaresearch.org/initiatives/dataset/

[7] AFEL online and social learning: http://data.afel-project.eu/catalogue/learning-analytics-dataset-v1/

- Use available long-term data sets to demonstrate validity of different algorithms and models for understanding factors associated with consumption and needs, as well as environmental impacts.

Because much data in this domain is not linked to consumers or individuals, a number of sources exist for public data sets. Here we focus on a small fraction of larger, public data sources in this space. Through the government's open data initiative,[8] researchers can get access to hundreds of agriculture-related databases and data sets including soil survey databases and maps, various food and crop databases, and local pollution data. DataRefuge has hundreds of climate, clean water, and pollution data sets.[9] Another publicly available, curated data set in this domain is the building performance database.[10] It is the largest database of information containing energy-related characteristics of commercial and residential buildings. The TRY database is a global archive of curated plant traits.[11] Hundreds of long-term ecological research data sets (habitat, animal population, environmental event data, etc.) are curated as part of the LTER research network.[12] Finally, IRIS is a research project that manages access to global earth science data, including earthquake, atmospheric, infrasonic, and hydrological data.[13]

3.3 Health

Health is a very broad domain with application areas in precision medicine, epidemics, health care cost management, and medication therapy and dosages, to name a few. There have been a number of recent successes in this domain including over 90% of hospitals and clinics using electronic healthcare records (EHR) and use these data for medication surveillance, the use of mobile health for real-time interventions, and the use of analytics generated from wearable devices for improving chronic disease patient care.

Over the next three years, promising directions for advancing the state of the art include:

- Improve methodologies and scale of causal inference techniques. While we have seen a large number of advances in machine learning, for precision health, mechanistic understanding is important.

- Develop methods that can be easily explained and interpreted by clinicians.

- With the increased use of EHR, use these data for more extensive public health understanding.

[8] Open Data Initiative: `https://www.data.gov`

[9] DataRefuge: `https://www.datarefuge.org/`

[10] Building Performance: `https://bpd.lbl.gov/`

[11] TRY (curated plant traits): `https://www.try-db.org`

[12] LTER: `https://portal.lternet.edu/nis/home.jsp`

[13] IRIS: `http://www.iris.edu/hq/`

- Use available data sets to improve patient diagnosis and identify precursors of illnesses earlier.
- As the number of mobile and wearable devices increases in this space, develop standardized ontologies and schemas for more rapid development of analytics-based healthcare outcomes.

Over the years, the federal government has helped fund a large number of studies that generated data. Many of these databases and data sets are accessible from the U.S. National Library of Medicine.[14] This site also links to other non-federal and state data repositories. Types of data sets include: Medicare provider utilization and payment data, health care outcome databases, inpatient hospital stays of children, healthcare claims data, CDC epidemiology data sets, emergency response data, veterans data, data on aging, substance abuse, etc. The National Library of Medicine also links to the HealthData.gov initiative that provides access to healthcare data sets from different Federal agencies. The Big Cities Health Coalition maintains aggregate health data from 28 large cities including data related to opioids, obesity, and tobacco.[15] The Healthcare Cost and Utilization Project (HCUP) developed through a Federal-State-Industry partnership has the largest collection of longitudinal hospital care data in the United States.[16] Finally, a number of bioinformatics databases have large data sets, including GenBank, a data repository of known genetic sequences, from the National Center for Biotechnology Information (NCBI)[17] and UniProt[18], a repository for protein data that combines data from different sources, e.g., SwissProt and PIR.

3.4 Policy

Big data has a number of different roles to play with regards to policy. First, big data can be used as evidence to inform policy and decision making. These data can be used to improve accountability through generated policy. Big data algorithms and data collection methods are also candidates for regulation and new technology-related policy, e.g., privacy and ethics related to using big data for different types of inference. Some recent successes in this area include the use of satellite data by NASA to improve food security through spatio-temporal data analysis, improved response to the Nepalese earthquake using opaque building images collected by drones, and updated water management policies in the Chesapeake Bay based on climate change and pollution models.

Over the next three years, promising directions for advancing the state of the art include:

- Develop case studies and automated detectors of potential misuse of big data to support specific policy agendas – identifying these types of sce-

[14] National Library of Medicine: https://www.nlm.nih.gov/

[15] Big Cities Health Coalition: http://www.bigcitieshealth.org/city-data/

[16] HCUP: https://hcup-us.ahrq.gov/databases.jsp

[17] NCBI: https://www.ncbi.nlm.nih.gov/genbank/

[18] UniProt: http://www.uniprot.org/

narios and educating policy makers and the public about them may help reduce this form of misuse.

- Incorporate explanations of error into analyses that use big data – develop standards and techniques for sharing levels of noise, bias, missing values, etc. to enable clearer communication of the big data results accompanying policy recommendations.

- Make models interpretable by policy makers so that big data can be used more readily in evidence-based policy recommendations.

Data in this domain are more scattered and very issue specific. For example, environment, energy, economic, finance and health are all examples of domains with data that may impact policy. Our focus on these domains should not be interpreted as being more important than other domains. Instead, they should be interpreted as domains that meeting participants were interested in discussing. There is no central data repository for public policy data sets. General population statistics, demographics, and voting data sets can be found at the Census Bureau.[19] Some other policy issues that have available data sets include: urban policy from the Urban Center for Computation and Data (UrbanCCD),[20] women and public policy including political participation, health, work and family, and safety from the Institute for Women's Policy Research (IWPR),[21] global health,[22][23] immigration,[24] and income inequity by the Organization for Economic Co-operation and Development (OECD).[25] There are also a number of simulation data sets that can be used to influence future policy, including the SUMO simulation of urban mobility/traffic on roads.[26]

3.5 The Economy and Finance

Big data and big data analytics have been applied in both finance and economics. Hedge funds have been using them successfully as alternative data inputs, scraping 100s of millions of websites daily. Twitter data has been used to predict a number of economic indicators including unemployment – some of these predictions have mixed success. Other areas of success include: market manipulation detection by searching for anomalies in daily and tick trading stocks, financial entity profile construction from public regulatory filings (SEC and FDIC), and identification of emerging risks.

Over the next three years, promising directions for advancing the state of the art include:

[19] Census data: https://www.census.gov/data/datasets.All.html

[20] UrbanCCD: http://www.urbanccd.org/research-and-tools/

[21] IWRP: https://iwpr.org/

[22] Global Health Data Exchange: http://ghdx.healthdata.org/

[23] World Health Organization: http://www.who.int/gho/database/en/

[24] EU Data Portal: https://data.europa.eu/euodp/data/dataset/L0q3araJ0g9Dk3TXZWkJg

[25] OECD: https://data.europa.eu/euodp/data/dataset/L0q3araJ0g9Dk3TXZWkJg

[26] SUMO: http://sumo.dlr.de/index.html

- Systematic generation of synthetic data sets that are designed as a “challenge” similar to the NIST challenge.[27]
- Detection of manipulation in financial markets
- Prediction of wider set of economic indicators

There is no central repository for data in this domain. The NIST data challenge was mentioned above. Real time stock data is relatively easy to obtain online. Company SEC filings can be obtained from the SEC search engine.[28] There are over 11 million filings that can be accessed. Data.gov hosts a diverse collection of datasets.[29]Another type of data that can be purchased through different credit companies are anonymized credit reports.

4 Final Thoughts

This report has highlighted a number of immediate research directions and available data sets for five different application areas. One evident finding is that the impact of big data varies considerably depending on the domain. While important research directions exist across all the domains, this variability is partially due to the variability of curated data sets in different domains. To increase the pace of research innovation, more effort and funds need to be devoted to data curation and data plumbing. Even after data curation, much work is still needed to make big data and data science concepts more accessible to a broader community. Initiatives like the DataCore and workplace data science training are vital for broadening the community and integrating big data analysis techniques into research across domains. Finally, we as a community need to be honest about big data as a field – while we push the boundaries, we also need to explain the limitations, assumptions, and biases that may be present in different analyses and that may differ from what people in different disciplines are accustomed to. We need to pause and think about the ethical implications of using certain data sets and pause to make sure we are preserving privacy and promoting fairness when developing new methods and algorithms.

References

[1] L. Singh, D. Logston, S. Nusser, H. Wactlar. (2015) *NITRD BDSI-2015 Workshop Report: Spearheading Innovation in the Face of Massive Data.*

[2] L. Singh, A. Deshpande, W. Zhou, S. Aluru, M. Balazinska, G. Biswas, A. Ganguly, A. Johri, F. Liu, M. Mahoney, C. North, K. Olukotun, A. Singh,

[27] NIST Financial Entity Identification and Information Integration: `https://ir.nist.gov/dsfin/`

[28] SEC Company Filings Search: `https://www.sec.gov/edgar/searchedgar/companysearch.html`

[29] `https://www.data.gov/finance/`

A. Smith, S. Venkatasubramanian. (2016) *NSF BDPI-2016 Workshop Report.*