Taxonomy of NSF's "Big Data" Research Projects

DRAFT VERSION

by Jacob Metcalf / October 15, 2014
Produced for Council for Big Data, Ethics, and Society[1]

This project categorizes and analyzes projects about "big data" supported by the US National Science Foundation between 2012 and 2014. It covers technical projects awarded by the BIGDATA Program under the Computer & Information Sciences & Engineering Directorate and social, ethical and behavioral projects awarded by the Social, Behavioral and Economic Sciences Directorate.

There is obviously an explosive growth in research projects that propose to *use* big data resources and techniques to answer empirical questions. For the sake of clarity, this analysis is restricted to projects that raise methodological or theoretical questions *about* big data.

Technical Research

Technical research projects propose to solve technical problems that are limiting the effectiveness of big data techniques. These projects can be understood as striking a balance between domain-generality and domain-specificity, and all of these projects can be placed somewhere on a continuum between generality and specificity. As is arguably the case for basic science, all of these projects *could* of course produce results relevant to wide range of domains. Furthermore, even the most domain-general algorithm research needs a dataset to work with, even if it could be *any* dataset in principle. Nonetheless, some projects start with a domain general problem, others start with a domain specific problem.

Domain general algorithmic problems

Domain general projects are those that attempt solve basic algorithmic, mathematical or engineering problems that exist generically for big data practitioners. These projects may or may not specify that the general problem is being *tested* within a domain. Commonly the tested domain is described at very high level, i.e., "government," "commerce," "research institutes," or "education." At times the PIs will focus the Intellectual Merit section of their grant on the domain general problems, and discuss the domain-specific uses as a Broader Impact (see example 1, see example 2). On its face that is an odd situation as one would not normally include basic research questions within the Broader Impact Criteria. However, it is indicative of how

[1] Funding for this Council was provided by the National Science Foundation (#IIS-1413864).

some data scientists view their relationship to particular datasets—they are secondary to the primary mathematical and engineering problems. In other words, new scientific knowledge is conceived as a consequence of big data techniques. Notably, this would appear to deviate substantially from the original intent of the Broader Impacts criterion.

Areas of concern for domain-general algorithmic research: When domain general algorithmic problems are tested on specific datasets how well will the solutions travel? What ethical artifacts are carried with them when they travel? Should funding agencies require more attention to the provenance of the data used to test domain general algorithms, even though the particular data used is ostensibly irrelevant? Do we really know that the datasets are irrelevant, or is that a uncritically assumed epistemic position?

Network: Connecting ever more devices and sensors to the Internet will create vast new sources of data. Making full use of the devices to achieve the many promised efficiency gains will require the ability to process and analyze that data, which will create new strains to networks. These projects seek to find ways of handling the increased network traffic generated by new torrents of data.

> Example: JetStream: A Flexible Distributed System for Online and In-Place Data Analysis. Michael Freedman et al. BIGDATA program.

Hardware: In some cases computer hardware is the limiting factor in the speed of large dataset analysis and storage, especially for research institutions that do not have an unlimited budget for hardware. Hardware limitations result in more risks of data corruption and human error. In other cases, contemporary algorithms often make use of legacy coding written for hardware in use 30 years ago and therefore do not make use of major hardware efficiencies available today (such as multiple processor 'cores' on a single chip). These projects attempt to improve the hardware or optimize algorithms to work with contemporary hardware.

> Example: An efficient, versatile, scalable, and portable storage system for scientific data containers. Werner Berger. BIGDATA program.

Scaling/scalability: A widespread problem is efficiently moving analytic tools between small datasets and large datasets—as the dataset gets larger, there is a substantial trade-off between accuracy/resolution of the analysis or query and the amount of computing resources used. These projects attempt to narrow or sidestep that trade-off.

> Example: Building a Mergeable and Interactive Distributed Data Layer for Big Data Summarization Systems. Feifei Li et al. BIDGATA program.

***Human labor expense/inefficiencies*:** Big data still relies on a surprising amount of human labor to collect and properly format datasets—getting to the analytic insights requires a lot of 'janitorial' labor. This preparatory labor is slow and expensive. These projects seek to lower the human time and cost of that preparation by automating it using machine learning techniques.

> Example: Big Data for Everyone. Tom Mitchell. BIGDATA Program

***Machine learning*:** Many of these projects involve improving machine learning algorithms. Improved machine learning capabilities reduce the number of decisions humans need to make when analyzing or using data, especially in the use of predictive algorithms. In scientific research, machine learning is especially important for sorting many-dimensional datasets into formats that can be interpreted by human researchers for producing knowledge.

> Example: Distribution-based machine learning for high dimensional datasets. Singh et al. BIGDATA Program.

Domain specific algorithmic problems

Domain specific projects focus on problems that have a limited scope, although they do often reference the possibly large scope of the proposed solutions.

Areas of concern for domain-specific algorithmic research: We might reasonably expect that these projects offer more specificity with regards to their datasets than the domain-general projects, however that is not always the case. Few directly address the provenance of their data as a technical or ethical question.

Genomics: Genomics has long been at the forefront of big data techniques in scientific research. Yet genomics has unique challenges because it relies on restrictive health records, draws on datasets that historically have had widely disparate formats and quality, deals with a patchwork of knowledge-fiefdoms at medical schools and makes use of scientific data locked up in journal articles (often in the big data-unfriendly .pdf format). Current research agendas focus on improving knowledge- and data-sharing (and therefore replicability of results), integrating data techniques with the much more powerful modern sequencers and creating the next generation of machine learning algorithms.

> Example: Genomes Galore - Core Techniques, Libraries, and Domain Specific Languages for High-Throughput DNA Sequencing. Alaru. BIGDATA program.

***Medical records*:** Like genomics, medical records represent some of the highest hopes and greatest challenges of big data techniques. Medical records are highly protected ethically and legally, but effective treatment requires sharing data and the health care system is highly fractured. Many records remain in written form or walled off behind proprietary systems. If legal and technological hurdles are overcome, big data techniques may enable predictive algorithms that improve preventative care and propagate best treatment practices.

> Example: Patient-level predictive modeling from massive longitudinal databases. Suchard et al. BIGDATA program.

Geospatial/Mapping: Complex, information-rich maps draw on data from many different, often-noisy, sources. Much of the mapping data currently available is stored as pixels, which is a format that is not particularly easy to connect to other sources of data inside 3D models. Pulling together these datasets and representing them as a coherent visual entity for users is a significant algorithmic challenge.

Example: Semantic Modeling of Cities from Scanned Data. Thomas Funkhouser. BIGDATA program.

Scientific collaboration/knowledge sharing: The sheer quantity of data makes it hard for scientists and engineers to find and utilize relevant data collected by others. These projects develop new modes for sharing, discovering, visualizing and indexing shared research data in ways that facilitate collaboration.

Example: Coupling Data-Intensive Modeling, Simulation, and Visualization with Human Facilities for Design: Applications to Next-Generation Medical Device Prototyping. Daniel Keefe et al. BIGDATA program.

Linguistics: Translation is an interesting opportunity and challenge for big data techniques. Typologically distinct languages do not easily map onto each other and yet there is an ever increasing demand to translate between languages used in far corners of the globe. Machine learning continues to improve automatic translation of natural languages but can mangle even typologically similar languages. The Internet has also now presented a new wealth of widely distributed sample data sets of natural language use, particularly on social media platforms.

Example: Big Multilinguality for Data-Driven Lexical Semantics. Smith et al. BIGDATA program.

Internet of things: The proliferation of internet-connected devices will dramatically increase the quantities of data produced, eventually sent over networks, and analyzed. This will introduce perhaps unexpected strains on the systems at the core of big data analytics, such a network bandwidth and storage. Although networked devices like smart appliances may not be as data-heavy as mobile phones, their proliferation will present new problems, which big data will need to adapt to.

Example: Real Time Observation Analysis for Healthcare Applications via Automatic Adaptation to Hardware Limitations. Hauptmann. BIGDATA program.

Social and ethical research

The NSF also funds an increasing number of workshops and projects (including dissertation awards and postdocs) directed at the social and ethical aspects of big data. A notable trend in these grants is the emphasis on agenda setting—many of these grants specifically note the lack of an established consensus about how humanists and social scientists should approach big data. Therefore many of these grants are structured around fostering dialog across disciplines about the consequences of big data for society, knowledge, power, and ethics. There is also concerted effort to figure out what sort of object "big data" is and situate it inside the methods and analytics of science studies. How does it travel? What are its effects outside of databases? How does it recruit actors?

Areas of concern for social and ethical research: Surprisingly, there are almost as many SBE grants *about* big data as there are grants awarded by the BIGDATA program to build actual big data tools. Topics in the humanities and social sciences that explode in popularity can flare

and fade out quickly. What will it take to build a sustainable, well-grounded research program about the social and ethical dynamics of big data? What can researchers in these areas do to productively interact with working on technical projects and leverage insights to get ahead of major social and ethical problems resulting from big data?

Privacy: These projects have a somewhat ambivalent attitude about privacy. Privacy appears to be a conversation-ender for social scientists, humanists and engineers, but it is a concept that does not easily fade away. There is a widespread consensus that big data is fundamentally altering the meaning of privacy for everyone involved.

> Example: Workshop: Privacy In An Era Of Big Data-Temple University; May 20-21, 2015. Pavlou. SECURE &TRUSTWORTHY CYBERSPACE program.

***Power/Persons/Society*:** Big data has the potential to engender substantial changes to the ways that we think of power, personhood and social institutions. Big data rhetoric promises to individualize everything, but doing so requires that aspects of our persons that have long been considered private be consolidated in datasets and run through algorithmic comparisons with untold numbers of other persons. This process redistributes power to institutions that are capable of leveraging big data techniques and creates pockets of resistance among those who want to protect established norms. There are a number of research projects and workshops that propose to grapple with this dynamic under a range big data topics, such as genomics and consumer privacy.

> Example: Open Data/Private Persons: Forging a New Social Contract for Biomedicine in an Age of Genomics and Big Data. Reardon. SCIENCE, TECH & SOCIETY program.

***Scientific Collaboration/Networking*:** Big data basic science research provides social scientists a new field of knowledge production to study as an ethnographic site. Scientific work and collaboration in big data fields, such as genomics or cosmology, are often far-flung enterprises that have no 'lab' in the traditional sense of the word. These projects are seeking modes of analysis that can track work and knowledge in this changing landscape.

> Example: Discovering Collaboration Network Structures and Dynamics in Big Data. Qin. SCIENCE OF SCIENCE POLICY program.

Governance/Civil Society: Big data could have substantial impact on matters of governance. Data collected on populations can inform policy and make services more efficient. However, where big data is leveraged as a tool to understand and serve populations it also lends itself to surveillance and abuse. One the other side of the coin, big data resources also provide new tools to journalists and civil society actors monitoring how governments behave and how political actors/networks drive changes in governance.

> Example: The Power of Policy Ideas: Tracing Language in Policymaking. Wilkerson. Political Science program.

Data infrastructure/archives: As more social science and humanities research questions rely on big data techniques themselves there is increasing demand for archiving and sharing data. The

NSF is funding a number of projects that build data infrastructures, standards and storage for research data and metadata collected in these fields.

> Example: Center for Historical Information and Analysis (CHIA). Manning. DATANET program.

***Economics*:** The proliferation of opportunities to collect real-time data on economic choices in many different markets has changed economics research substantially. Researchers are able to gain a finer and broader understanding of how incentives drive individual behavior and how data about these choices are in turn shaping innovation, investment and marketing. The NSF is funding a variety of projects examining methodological and empirical questions arising from this new resource.

> Example: Economics of Internet Markets. Levin. ECONOMICS program.