Data Driven Discovery In the Social, Behavioral, and Economic Sciences

Simon Appleford, Marshall Scott Poole, Kevin Franklin, Peter Bajcsy, Alan B. Craig,
Institute for Computing in the Humanities, Arts, and Social Science
University of Illinois Urbana-Champaign

Please Send Correspondence to:
sapplefo@ncsa.uiuc.edu

Theory Building and Data Driven Discovery

Simon Appleford, Marshall Scott Poole, Kevin Franklin, Peter Bajcsy, Alan B. Craig,
Institute for Computing in the Humanities, Arts, and Social Science
University of Illinois Urbana-Champaign

Abstract

With the advent of petascale computing, the social, behavioral, and economic (SBE) sciences have an opportunity to transform their research and move away from the limits imposed upon them by the current state of the art in technology. Central to this paradigmatic shift, however, is the need for SBE researchers to confront the Grand Challenge question, *How can the social, behavioral, and economic sciences harness vast stores of digital data for scientific inquiry?* This white paper proposes the development of a new type of inquiry, Data Driven Discovery (D3), which will integrate the capacity to collect and analyze huge digital datasets, analytics to uncover patterns and relationships, and rigorous theory. Central to this program is the need to develop new models for how to manage and analyze the large amounts of data available to researchers. Advancements are proposed for acquiring data, analytics, and conduct of research to allow scholars to address complex issues that currently challenge research and that will allow them to make transformational discoveries in their research and scholarship.

Data Driven Discovery In the Social, Behavioral, and Economic Sciences

Simon Appleford, Marshall Scott Poole, Kevin Franklin, Peter Bajcsy, Alan B. Craig,
Institute for Computing in the Humanities, Arts, and Social Science
University of Illinois Urbana-Champaign

The coming decade presents an opportunity for the social, behavioral, and economic sciences to transform themselves in ways that parallel the transformation of the physical and biological sciences over the past two decades. Large scale, high performance computing has fundamentally altered inquiry in astronomy, genetics, and other sciences. Advances in high performance and data-intensive computing that now allow researchers to query huge digital datasets and use computational analytics will similarly offer the social, behavioral, and economic (SBE) disciplines the opportunity to conduct unprecedented inquiries into the nature of society, human behavior, and the economy.

The paradigmatic example of these possibilities is the advent of fMRI and other brain imaging technologies. These tools, which depend on complex computational and storage resources, have fundamentally altered how cognitive science is done. They are generating novel, transformational discoveries and stimulating new theories that illuminate the nature of cognition, the interrelationship between biology and cognition, and other critical phenomena. High performance and data-intensive computing hold similar potential to transform the paradigm by which SBE scholars conduct their research.

However, capturing the promise of these technological advances depends on answering a Grand Challenge, namely:

> **How can the social, behavioral, and economic sciences harness vast stores of digital data for scientific inquiry?**

Answering this challenge requires the development of a new type of inquiry, Data Driven Discovery (D3). Data Driven Discovery integrates the capacity to collect, manage, and analyze huge digital datasets, analytics to uncover patterns and relationships, and rigorous theory. D3 will open up new paths to discovery in the SBE sciences by capitalizing on an important well of discovery the physical and biological sciences have regularly tapped, a body of unexplained, yet dependable and reliable data that poses a puzzle, something that requires explanation.

The development of new instruments such as ever-better telescopes and gene sequencing technologies have resulted directly from the requirements of the physical and biological sciences to have ever better data. Theories guide the development of most new instruments. Once developed, however, instruments have an additional impact on inquiry: they generate observations that are not within the scope of current theory. These inexplicable observations have stimulated new discoveries and insights and have proven to be important drivers of science. The lack of error in these data (compared to that in current SBE data) and the sheer quantity of data make them uniquely valuable contributors to the progress of science.

The advent of digital technologies to document human communications and transactions, massive digital repositories of documents, images, and other human artifacts, and associated data and tools such as GIS systems, video banks, and data mining capabilities, afford SBE scholars with the ability to conduct research on a par with that in the physical and biological sciences. They offer the capacity to surface puzzling facts that challenge existing conceptualizations and theoretical frames. But these technologies themselves are challenging and in the developmental stages. And much important data is locked in a format that is difficult for digital instruments to extract and analyze.

Data

We are now entering an age where data is (or can be) abundant. In some cases, this data is already available and can be automatically collected, such as archives of email messages. In other cases, researchers must acquire or generate the data they need. Data handling is thus a grand challenge in its own right. Huge data sets from large multi-year surveys, email archives, sampling of internet sites over time, and video recordings all require high performance computing and data management. For example, monitoring the behavior of multiple emergency responders at a high degree of temporal and spatial fidelity requires the ability to integrate multiple streams of video, audio, GIS, and sensor data (Poole et al, 2010).

The first challenge is the sheer size of databases involved. The soon-to-be released 1940 Census, for example, will be made available to the public across 4,643 rolls of microfilm. To create a digital, archival, quality version of these records would result in 3.25 million images and 125.8TB of data that must then be processed to allow for the extraction of as much information from the documents as possible. Other types of data represent even more daunting challenges: the Bush White House generated an average of almost 250,000 email messages per *day*; twitter users send an average of 1.9 million tweets a day; and a single hour of uncompressed video data can require anything from 10 to 834GB of storage.

Another challenge is related to transforming data into formats suitable for digital manipulation. For example, currently there are 1% and 5% samples of the US census available in digital form, but any study requiring a more comprehensive volume of census data is limited by the need to digitize the records and perform OCR on hand written documents, a process that requires significant computational and human resources. The question of how to unlock data that is currently found in machine-readable format is a prime candidate for research.

There is also a need for better mechanisms for organizing and sharing data. Traditional monolithic databases should be supplemented by tools that facilitate exploration of relationships that aren't supported by the database schema. Mechanisms are needed to automatically extract metadata from the underlying datasets. Beyond simply identifying data such as file creation time, etc., the ability to automatically extract information such as that needed to geo-reference the document are needed. In order to facilitate sharing and reuse of data, as well as the ability to integrate different data sources, it is essential for data in the behavioral sciences to adhere to data standards and best practices. Additionally, for data sharing it is important to look at data models like RDF and semantic web

standards. In order to ensure reproducibility and to verify the history of the data, it is essential to track data provenance. Finally, in order to ensure the greatest opportunity to use the data in the future, there needs to be a mechanism for expediting the deposit of data into institutional repositories for long term preservation.

Analytics

The nature of D3 datasets necessitates the development of novel analytical methods specifically suited to the SBE sciences. Methods such as visualization and data/text/image mining are needed to help make sense of the data and to supplement more traditional statistical analysis. These approaches are truly in their infancy in terms of meaningful applications in SBE inquiry. For example, in studies of activity in online games, mapping social networks among participants is a logical way to visualize structure in the data. Physicists, computer scientists and mathematicians have developed methodologies to map and to characterize large scale networks using global parameters. Many of the questions posed by the SBE disciplines, however, require methods that focus more on local properties, such as local network structures, generation of linkages based on social or cognitive principles, and variables characterizing individual actors or units in terms of network properties, and methods to capture these variables must be developed.

Some D3 datasets, like the census, refer to one static slice of time. However, the nature of digital data collection in real time generates datasets that capture micro-level phenomena (key strokes, transactions, game plays, registrations, etc.) over time (often at the level of seconds or hundredths of seconds). The resulting detailed, longitudinal, large-sample datasets have not previously been common in SBE research. They offer SBE scholars the capability to conduct studies of dynamics in phenomena which have traditionally been studied only with cross-sectional designs. For example, with a four month download of game data it is possible to study samples of 2000 teams with shifting membership over multiple tasks for long periods of time. To be able to conduct such studies it is necessary to develop methods for identifying more macro-level patterns (e.g. team behavior) from micro-level records (e.g., records of individual player behavior) and for analyzing longitudinal processes. Techniques for transforming microlevel data into indicators of more macro constructs are largely ad hoc and need more formal development. With the exception of time series and event history methods, methodologies for analysis of longitudinal processes are also underdeveloped for the types of phenomena commonly studied in SBE inquiry.

These large sample, often longitudinal datasets offer reference points for large-scale simulation and mathematical modeling efforts. SBE scholars can use models to generate predictions which can be compared to the data. Modeling efforts can also draw on datasets for parameters or starting conditions. Development of such large-scale modeling efforts are likely to require supercomputing or grid computing support.

To take advantage of these resources we must develop a national infrastructure that supports the needs of SBE researchers. Current technical and policy infrastructures do not meet the needs of many projects in SBE computing because they are primarily suited to batch operations. SBE researchers, however,

often require a different mode of operation whereby their data is hosted in a large database that needs to be queried and analyzed in real time. This poses quite a different problem than batch operations and requires different computing infrastructure and in some cases supercomputer-scale computational capacity.

Theory

D3 will change the nature of theory development in the SBE sciences. Rather than starting with top down theory only, D3 often poses puzzles in the form of unexplained patterns (both static and dynamic). Explaining these patterns using existing theory and models requires retroductive reasoning, described by Charles Pierce (1955) as abduction, in which one or more theories or models that might account for observed patterns are evaluated against the patterns and other relevant data. Protocols for retroduction that protect against capitalizing on chance and involve multiple cycles from model to data and back need to be developed. The large samples typical of D3 datasets also allow retroduction of a model on one subset of the data and then validation of the model on another.

The use of mathematical and simulation modeling will also promote more precise specification of SBE theories. Verbal theoretical formulations often are more ambiguous than formal representations in modeling and mathematical terms, and the capability of D3 inquiry to support formal modeling can help in specifying theories. Development of dynamic theories will also be greatly facilitated by model-based theorization. On the other hand, difficulties in moving from verbal to modeling formulations can stimulate novel forms and approaches to modeling

Conduct of Research

The huge databases generated by D3 research offer opportunities for the development of long-term, distributed scholarly communities analogous to those organized around telescopes in astronomy. SBE scientists are accustomed to team approaches, but most teams are small and not up to the task of analyzing huge datastores over multi-year projects. Consequently, it will be necessary to develop a comprehensive program to address the training and education of a new breed of SBE researcher. In much the same way that focused training programs were created for scientists, and educational programs surrounding computational science were developed, a similar environment must be created for the SBE community. Additionally, a resource to aid in application development, and to actually carry out the application development for projects deemed of high national importance should be made available throughout the social science research community.

Conclusion

By harnessing the vast stores of digital data for scientific inquiry and using newly available computational techniques such as advanced data acquisition, data storage and management, user-friendly data mining and visualization technologies, large-scale modeling and simulation, massive text and visual searches with complex relational analysis, the SBE sciences are poised to make a paradigmatic transformation in how they conduct their research, the scale of their inquiries, and the scope of SBE discovery. The advent of petascale computing will allow SBE scientists to move beyond the constraints

imposed on them by the current state of the art to address complex issues that currently challenge SBE research. NSF should position itself to catalyze the D3 transformation in the SBE sciences.

References

Poole, M. S., Bajcsy, P., Contractor, N., Espelage, D., Forsyth, D., Hasegawa-Johnson, M., and Pena-Mora, F. (2010). GroupScope: Instrumenting Research on Interaction Networks in Complex Social Contexts. University of Illinois Urbana-Champaign. Project funded by the CDI Program of NSF.

Peirce, C. S. (1955). *Philosophical Writings of Peirce* ed. J. Buchler. New York: Dover.