RPR Review of Policy Research

Policy Studies Organization PSO

The Promise and Potential of Big Data: A Case for Discovery Informatics

Vasant G. Honavar
College of Information Sciences and Technology, Pennsylvania State University

Abstract

The emergence of "big data" offers unprecedented opportunities for not only accelerating scientific advances, but also enabling new modes of discovery. While we understand how to automate routine aspects of data management and analytics, most elements of the scientific process currently require considerable human expertise and effort. We argue that realizing the full potential of data to accelerate discovery calls for a concerted effort in advancing Discovery Informatics: (i) understanding, formalization, and information processing descriptions of the entire scientific process; (ii) design, development, and evaluation of the computational artifacts (representations and processes) that embody such understanding; and (iii) application of the resulting artifacts and systems to advance science (by augmenting individual or collective human efforts, or by fully automating science).

KEY WORDS: innovation, ICTs, health & medicine, biotechnology, Big Data, discovery, Discovery Informatics

The Transformative Potential of Big Data

Rapid advances in instrumentation and sensors, digital storage, computing, and communications have resulted in a transformation of many historically data-poor sciences into increasingly data-rich sciences. New discoveries in biological, physical, cognitive, and social sciences, and engineering are increasingly being driven by our ability to acquire, share, integrate and analyze, and build predictive models from data. Modern data-analytics techniques that integrate sophisticated probabilistic models, statistical inference, and scalable data structures and algorithms into powerful machine-learning algorithms have resulted in powerful ways to extract actionable knowledge from data in virtually every area of human endeavor. Creative applications of data analytics are enabling biologists to gain insights into how living systems acquire, encode, process, and transmit information; neuroscientists to uncover the neural bases of cognition; health scientists to not only diagnose and treat diseases but also help individuals make healthy choices; economists to understand markets; security analysts to uncover threats to national security; social scientists to study the evolution and dynamics of social networks; and scholars to gain new understandings of literature, arts, history, and cultures through advances in the digital humanities.

The exponential growth in the volume, variety, and velocity of data in virtually every area of human endeavor has led to the emergence of "big data." The current state of affairs in biomedical sciences helps illustrate the impact of "big data." In 2011, the number of peer-reviewed biomedical research articles appearing in PubMed exceeded 2,700 per day. A consequence of this explosion in the rate of growth of scientific literature is that it is virtually impossible for a scientist to keep up with all of the findings in his or her discipline. Peer-reviewed research articles constitute only one kind of big data. Many scientific investigations increasingly need to draw on

Review of Policy Research, Volume 31, Number 4 (2014) 10.1111/ropr.12080

experimental and observational data, expertise, and results from multiple disparate data sources across multiple disciplines. Consequently, there is a huge, rapidly widening, gap between our ability to accumulate big data and our ability to make effective use of such data to advance discovery. Hence, there is an urgent need for sophisticated tools for data curation, management, and analytics, including scalable tools for storing, indexing, annotating, retrieving, organizing, assessing the reliability of, and analyzing data (including observational and experimental data, literature, images, spatial, temporal, richly structured, e.g., network data).

Realizing the Potential of Big Data: The Leap from "Stamp Collecting" to "Physics"

"All science is either stamp collecting or physics," said Rutherford. The emergence of big data has led to an exponential growth in the volume and variety, and rate of acquisition of stamp collections. Advances in computing, storage, and communication technologies have made it possible to organize, annotate, link, share, discuss, and analyze increasingly voluminous, exquisitely diverse data, i.e., "stamp collections." Yet our current understanding of complex biological, cognitive, economic, and social phenomena remains, much like the understanding of physics before Newton, descriptive, or in Rutherford's terminology, stamp collecting. What would it take for these disciplines to make the leap from stamp collecting to physics, i.e., from descriptive sciences to predictive sciences?

Realizing the Potential of Big Data: The Importance of Models

It was the invention of calculus by Newton and Leibnitz that for the first time allowed precise descriptions of rate of change, and hence fundamental constructs of classical physics such as velocity and acceleration, that helped transform the study of the physical universe from "stamp collecting" to "physics," from a descriptive science into a predictive science. It might be arguable as to whether it is possible to discover the analogs of simple laws like Newton's laws of motion that provide accurate predictive models of complex biological, cognitive, economic, and social phenomena. Fortunately however, the invention of the formal notion of computation as the process of manipulating symbolic descriptions by Turing and others, offers a powerful machinery, analogous to calculus for physics, for specifying precise recipes—algorithms—that can be used to describe the relationships between, and the processes that operate on, the biological, cognitive, economic, and social entities that make up the world around us. Because *anything* that is describable can be described using a computer program, algorithms provide a powerful substrate for specifying, and reasoning about, theories of the world. If we take this view, we understand a phenomenon when we have an algorithm that describes it at the desired level of abstraction. Thus, we will have a theory of protein folding when we can specify an algorithm that takes as input, a linear sequence of amino acids that make up the protein (and the relevant features of the cellular environment in which folding is to occur), and produces as output, a description of the three-dimensional structure of the protein (or more precisely, a set of stable configurations).

Making sense of data requires advances in knowledge representation languages and modeling formalisms for describing and predicting the underlying phenomena at varying levels of abstraction. A shift in emphasis from collecting and cataloging data to understanding the behavior of complex systems, i.e., from "stamp collecting" to "physics" calls for representation and modeling languages with precise formal semantics for describing, sharing, and communicating scientific models, theories, and hypotheses. The need for automation dictates that the models must be specified in a form that can be processed by computers; and queries against the model and data be translated into precise computational problems.

Realizing the Potential of Big Data: The Importance of Processes and Tools

While we understand how to automate routine aspects of data management and analytics, humans are still largely responsible for most elements of the scientific process (see Chalmers, 1999; Hacking, 1983; Rosenberg, 2000, for characterizations of the scientific process). Examples of elements of the scientific process that have largely resisted automation include: mapping the current state of knowledge; generating and prioritizing questions; designing studies; designing, prioritizing, planning, and executing experiments; interpreting results; forming hypotheses; drawing conclusions; replicating studies; validating claims; documenting studies; communicating results; reviewing results; and integrating results into the larger body of knowledge in a discipline. Hence, we need automated or interactive tools to support all of these key elements of the scientific process. Because science is increasingly a collaborative endeavour, we need sharable and communicable representations and processes, organizational and social structures and processes that facilitate collaborative discovery, including mechanisms for decomposing tasks, assigning tasks to and incentivizing participants, sharing relevant "mental models," combining results, and at least in some domains, engaging large numbers of participants with varying levels of expertise and ability in discovery ("citizen science").

Realizing the Potential of Big Data: The Informatics of Discovery

As our ability to gather digital information of all kinds outstrips our cognitive ability to process, assimilate, and use the information, realizing the potential of data—big and small—to extract useful knowledge to inform our decisions and actions and to make the leap from "stamp collecting" to "physics" in biological, cognitive, and social sciences, calls for deeper understanding of the processes of discovery and the methods and tools that embody such understanding to help accelerate discovery.

Realizing the transformative potential of data requires frameworks that organize the hypotheses that are under consideration, the data that supports them, the models that have been created from the data, and the hypotheses resulting from the models. Note that the processes of discovery have to do primarily with acquiring, organizing, verifying, validating, integrating, analyzing, reasoning with, and communicating information (models, hypotheses, theories, and explanations) concerning natural and built systems. Hence, computing, the science of information processing, offers not only a powerful formal framework and exploratory apparatus for sciences (Djorgovski, 2005) but also the theoretical and experimental tools for the study of the feasibility, structure, expression, and automation of the processes that underlie discovery.

While automating aspects of scientific discovery has been a topic of considerable interest in artificial intelligence (de Jong & Rip, 1997; Dzeroski & Todorovski, 2007; Glymour, 2004; Langley, Simon, Bradshaw, & Zytkow, 1987; Pearl, 2003; Shrager & Langley, 1990; Valdez-Perez, 1999), information science (Smalheiser, 2012; Swanson & Smalheiser, 1997), and cognitive science (Klahr, 2000), it is only relatively recently that some of the prerequisites for automating discovery (such as technology for automating data acquisition, databases, and knowledge bases that capture the relevant background knowledge in specific disciplines, e.g., biological sciences, open access to large bodies of scientific literature, technologies for connecting resources and experts, and for constructing and sharing scientific workflows) have become available (Gil & Hirsh, 2012). King et al. (2009) have demonstrated the possibility of automating science by building a robot scientist capable of autonomously generating and testing hypotheses, in this instance, concerning the functional genomics of yeast (*Saccharomyces cerevisiae*).

These developments, together with the transformative potential of big data across many areas of science and even the humanities, strongly argue for a concerted effort to revisit the challenges of automating aspects of discovery as well as developing computational tools to augment human abilities in the domain of scientific discovery. Realizing the full potential of big data to advance discovery calls for a new discipline, Discovery Informatics (Honavar, 2013), that aims to: understand and formalize the representations, processes, and organizational structures that are crucial to discovery in the sciences as well as the humanities; design, develop, and evaluate the computing and information artifacts that embody such understanding; and apply the resulting artifacts and systems to facilitate discovery.

Conclusion

The emergence of “big data” offers unprecedented opportunities for not only accelerating scientific advances, but also enabling new modes of discovery. While we understand how to automate routine aspects of data management and analytics, most elements of the scientific process currently require considerable human expertise and effort and have resisted automation. We have argued that a concerted effort to advance the informatics of discovery is of utmost importance in the success of efforts to realize the full transformative potential of big data.

Advances in Discovery Informatics inevitably require collaborative projects that bring together bench scientists in one or more specific domains of inquiry, e.g., the biomedical sciences, with information and computer scientists, organizational and social scientists, cognitive scientists, philosophers of science, to study and formalize the representations, processes, and organizational structures that are crucial to discovery. Such collaborations would be hard to sustain in the absence of funding mechanisms that support not only collaborative research in Discovery Informatics but also fundamental research in the relevant disciplines including computer science, informatics, artificial intelligence, robotics, data and computing infrastructure, cognitive science and social science on the one hand and the applications of discovery informatics in specific domains that are ripe for such efforts, e.g., systems biology, materials science, health sciences, behavior and brain sciences on the other. Particularly important are funding mechanisms that support interdisciplinary research-based pre- and postdoctoral training opportunities for preparing a diverse cadre of young scientists to pursue careers in Discovery Informatics.

Given the critical role of Discovery Informatics in realizing the transformative potential of big data investments in Discovery Informatics are likely to directly benefit multiple areas of national priority including education, food, health, environment, energy, and security.

About the Author

Vasant G. Honavar is a Professor and Edward Frymoyer Chair of Information Sciences and Technology and Professor of Bioinformatics and Genomics and of Neuroscience at Pennsylvania State University where he currently leads the Artificial Intelligence Research Laboratory and a research initiative in Discovery Informatics. Honavar has served as a Program Director in the Information and Intelligent Systems Division at the National Science Foundation (during 2010–2013) where he contributed to multiple programs including Information Integration and Informatics, Smart and Connected Health, and led the Big Data Science and Engineering Program. His research has been published in over 250 peer-reviewed publications and he currently serves on the editorial boards of several journals and is a general co-chair of the IEEE International Conference on Big Data (2014).

References

Chalmers, A. F. (1999). *What is this thing called science?* Queensland, Australia: University of Queensland Press.

de Jong, H., & Rip, A. (1997). The computer revolution in science: Steps towards the realization of computer-supported discovery environments. *Artificial Intelligence*, *91*, 225–256.

Djorgovski. (2005). Virtual astronomy, information technology, and the new scientific methodology. In V. Di Gesu & D. Tegolo (Eds.), *Proc. of CAMP05, "computer architectures for machine perception"* (pp. 125–132). New York, NY: IEEE Press.

Dzeroski, S., & Todorovski, L. (Eds.). (2007). *Computational discovery of communicable scientific knowledge*. Berlin, Germany: Springer.

Gil, Y., & Hirsh, Y. (2012). Discovery informatics: AI opportunities in scientific discovery. AAAI Technical Report FS-12-03.

Glymour, C. (2004). The automation of discovery. *Daedelus*, *Winter*, 69–77.

Hacking, I. (1983). *Representing and intervening. Introductory topics in the philosophy of science*. Cambridge, England: Cambridge University Press.

Honavar, V. (2013). *From Data Analytics to Discovery Informatics*. In: Data Science: Unlocking the Power of Big Data. National Institutes of Health Videocast. Retrieved from http://videocast.nih.gov/summary.asp?bhjs=0&File=17798

King, R. D., Rowland, J., Oliver, S. G., Young, M., Aubrey, W., Byrne, E., et al. (2009). The automation of science. *Science*, *324*(5923), 85–89.

Klahr, D. (2000). *Exploring science: The cognition and development of discovery processes*. Cambridge, MA: MIT Press.

Langley, P., Simon, H. A., Bradshaw, G. L., & Zytkow, J. M. (1987). *Scientific discovery: Computational explorations of the creative processes*. Cambridge, MA: MIT Press.

Pearl, J. (2003). *Causality: Models, reasoning, and inference*. Cambridge, England: Cambridge University Press.

Rosenberg, A. (2000). *Philosophy of science*. London, England: Routledge Press.

Shrager, J., & Langley, P. (Eds.). (1990). *Computational models of scientific discovery and theory formation*. San Mateo, CA: Morgan Kaufmann.

Smalheiser, N. R. (2012). Literature-based discovery: Beyond the ABCs. *Journal of the American Society for Information Science. American Society for Information Science*, *63*, 218–224. doi:10.1002/asi.21599

Swanson, D. R., & Smalheiser, N. R. (1997). An interactive system for finding complementary literatures: A stimulus to scientific discovery. *Artificial Intelligence*, *91*, 183–203.

Valdez-Perez, R. E. (1999). Principles of human-computer collaboration for knowledge discovery in science. *Artificial Intelligence*, *107*, 335–346.