

ACCELERATING SCIENCE: A COMPUTING RESEARCH AGENDA

Vasant G. Honavar¹, Mark D. Hill², and Katherine Yelick³

Computing Community Consortium

2/19/2016

Version 1

Abstract

The emergence of “big data” offers unprecedented opportunities for not only accelerating scientific advances but also enabling new modes of discovery. Scientific progress in many disciplines is increasingly enabled by our ability to examine natural phenomena through the computational lens, i.e., using algorithmic or information processing abstractions of the underlying processes; and our ability to acquire, share, integrate and analyze disparate types of data. However, there is a huge gap between our ability to acquire, store, and process data and our ability to make effective use of the data to advance discovery. Despite successful automation of routine aspects of data management and analytics, most elements of the scientific process currently require considerable human expertise and effort. Accelerating science to keep pace with the rate of data acquisition and data processing calls for the development of algorithmic or information processing abstractions, coupled with formal methods and tools for modeling and simulation of natural processes as well as major innovations in *cognitive tools* for scientists, i.e., computational tools that leverage and extend the reach of human intellect, and partner with humans on a broad range of tasks in scientific discovery (e.g., identifying, prioritizing formulating questions, designing, prioritizing and executing experiments designed to answer a chosen question, drawing inferences and evaluating the results, and formulating new questions, in a closed-loop fashion). This calls for concerted research agenda aimed at: Development, analysis, integration, sharing, and simulation of algorithmic or information processing abstractions of natural processes, coupled with formal methods and tools for their analyses and simulation; Innovations in cognitive tools that augment and extend human intellect and partner with humans in all aspects of science. This in turn requires: the formalization, development, analysis, of algorithmic or information processing abstractions of various aspects of the scientific process; the development of computational artifacts (representations, processes, protocols, workflows, software) that embody such understanding; and the integration of the resulting cognitive tools into collaborative human-machine systems and infrastructure to advance science.

OVERVIEW

Tycho Brahe gathered considerable and accurate data on the movement of the planets (“big data” for his time). However, this data did not find real value until Johannes Kepler used it to discover his three

¹ Professor and Edward Frymoyer Chair, College of Information Sciences and Technology, Professor of Computer Science, Bioinformatics and Computational Biology and Neuroscience Graduate Programs, Director, Center for Big Data Analytics and Discovery Informatics, and Associate Director, Institute for Cyberscience, Pennsylvania State University, University Park, PA 16802. Email: vhonavar@ist.psu.edu; Web: <http://faculty.ist.psu.edu/vhonavar>

² John P. Morgridge and Gene P. Amdahl Professor and Chair, Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI 53706. Email: markhill@cs.wisc.edu; Web: <http://pages.cs.wisc.edu/~markhill/>

³ Professor, Computer Science Division, University of California at Berkeley, Berkeley, CA 94720 and Associate Laboratory Director of Computing Sciences, Lawrence Berkeley National Laboratory. Email: yelick@berkeley.edu; Web: <http://www.cs.berkeley.edu/~yelick/>

laws of planetary motion. Later Isaac Newton used these laws and other data to derive his unified laws of motion, and lay the foundations of classical physics. To do so, he had to invent calculus for describing such things as rates of change. Brahe, Kepler, and Newton were all engaged in the practice of science, a systematic process for acquiring knowledge through observation or experimentation and developing theories to describe and explain natural phenomena. The past centuries have witnessed major scientific breakthroughs as a result of advances in instruments of observation, formalisms for describing the laws of nature, and improved tools for calculation.

Today, the experimental instruments are more powerful, the scientific questions more complex, and the mathematical, statistical and computational methods for analyzing data have become more sophisticated. The resulting emergence of “big data” offers unprecedented opportunities for accelerating science. Arguably, “big data” accelerates Brahe’s part of the scientific endeavor, and increasingly, Kepler’s part, with the increasing use of machine learning for building models from data. Nevertheless, many aspects of the scientific process (designing, prioritizing and executing experiments, organizing data, integrating data, identifying patterns, drawing inferences and interpreting results) constitute an even greater bottleneck than ever.

The goal of this white paper is to articulate a research agenda for developing cognitive tools that can augment human intellect, and partner with humans on all aspects of the scientific process, including in particular, those that are exacerbated by “big data.” We argue that there is great opportunity for dramatically accelerating science and enabling new modes of scientific discovery, perhaps even empowering and enabling the future likes of Kepler or Newton in the era of big data.

The benefits of accelerating science extend well beyond the scientific community to all of humanity. Imagine: Precision health regimens that take into account not only one’s genetic makeup, but also environment, and lifestyle; Personalized education that optimizes curriculum, pedagogy, etc. to optimize the learning outcomes for each individual; Precision agriculture that optimizes everything from the choice of crops to water and fertilizer use to optimize yield and impact on the environment. These are just a start, however, as in the 21st century we should be able to invent technologies undreamed of in this century’s early years, as who in 1900 could have anticipated 20th century advances, such as the Internet (no computers yet) or DNA sequencing (DNA structure unknown)?

ACCELERATING SCIENCE: THE VALUE PROPOSITION

Imagine a world in which scientists work with cognitive tools that can

- Given access to literature and data:
 - Create and share a knowledge base that summarizes what we know about a scientific question (annotated with uncertainty, provenance, and underlying assumptions);
 - Summarize and prioritize questions that need to be answered to achieve an overall scientific objective (e.g., understanding the molecular mechanisms that underlie cancer);
 - Identify and rank alternative explanations of an observation based on the current state of scientific understanding in a given field;
 - Design and prioritize study techniques;
 - Construct a computational model, e.g., a network of genes that orchestrate a specific biological process of interest, that make experimentally testable predictions.
- Given a conjecture:
 - Identify data that support or refute the conjecture;
 - Identify simulations that can interpret the theory, e.g., over time or in various physical settings;
 - Design and prioritize, orchestrate, and execute experiments.
- Given an experimental design, experimental results, and access to literature:

- Create a plan for replicating the study and validating the claims;
- Generate and rank alternative interpretations of the data;
- Document the study, communicate results;
- Integrate results into the larger body of knowledge within or across disciplines.
- Given a collection of experimental and observational studies:
 - Infer a causal effect of interest, e.g., the role of a specific gene or combination of genes in a specific biological process;
 - Calculate scientific parameters, e.g., geophysical characteristics affecting earthquakes, by solving an inverse problem by comparing simulations to the observations.
- Given a scientific question and a network of researchers, assemble a team that is best equipped to answer the question;
- Track scientific progress, evolution of scientific disciplines, and scientific impact.

Cognitive tools for accelerating science could lead to dramatic increases in scientific productivity by increasing efficiency of the key steps in scientific process, and in the quality of science that is carried out (by reducing error, enhancing reproducibility), allow scientific treatment of topics that were previously impossible to address, and enable new modes of discovery that leverage large amounts of data, knowledge, and automated inference. The sections that follow attempt to further flesh out our vision for accelerating science by accelerating increasingly larger fractions of the scientific process.

THE SCIENTIFIC PROCESS

To understand where the major bottlenecks to scientific progress are, it is useful to revisit a simplified model or template of the scientific process (See Hacking, 1983; Chalmers, 1999; Rosenberg, 2000 for reviews). Figure 1 summarizes the key elements of the scientific process.

Typically, scientific inquiry starts with a question within a domain of study, e.g., biology. With the question in hand, one has to assemble the background information and acquire the data necessary to answer the question. Then

one proceeds to construct one or more models from data (and background information). Choosing a small set of models from among a much larger set of candidates involves additional considerations (simplicity, consistency with what else is known), etc. The models can be used to advance hypotheses that result, ideally, in testable predictions. The observations or experiments designed to test the predictions yield additional data that feed into the larger scientific process. Science is a social endeavor, with multiple individuals and teams, driven by intrinsic as well as extrinsic incentives. Scientific findings go through peer review, communication, and publication, and replication before they are integrated into the larger body of knowledge in the relevant discipline.

It is worth noting that there is considerable variability across scientific disciplines, e.g., in cosmology, where there is little possibility of executing designed experiments, one typically has to make do with observational data or the results of ‘natural’ experiments. Nevertheless, it is clear that the processes of acquiring, organizing, verifying, validating, integrating, analyzing, reasoning with, and communicating

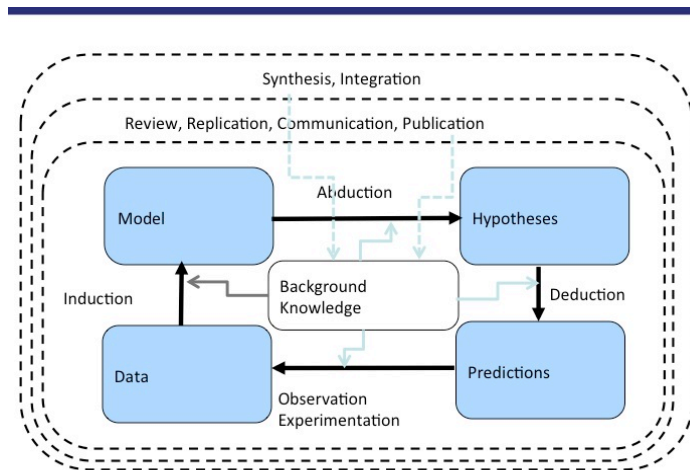


Figure 1: A Cartoon of the Scientific Process

information (models, hypotheses, theories, explanations) about natural and built systems lie at the heart of the scientific enterprise.

ACCELERATING SCIENCE: TRADITIONAL ENABLERS

Major scientific advances are often enabled by:

- Advances in the instruments of observation (new measurement devices or methods or making it possible to acquire data of new modalities, higher resolution in time or space, or in larger volumes than previously possible).
- Development of mathematical models and methods for representing and reasoning about scientific hypotheses and theories (e.g., the invention of calculus by Newton and Leibnitz that were necessary for the advances in physics);
- Development of effective tools for data analysis and simulation (e.g., the invention of the computer that enabled among other things, solution of systems of linear equations, simulation of complex models of physical, biological, and cognitive processes);
- Cross-fertilization and integration of concepts, experimental methods, data, tools, hypotheses, theories, across disciplines (e.g., the emergence of molecular biology through convergence of biological and physical sciences).

In what follows, we argue that the emergence of big data and the ability to examine natural processes using the computational lens (Karp, 2011), offer the possibility of rapid acceleration of science. However, realizing this requires algorithmic or information processing abstractions of natural processes, coupled with formal methods and tools for their analyses and simulation; cognitive tools that augment and extend human intellect and partner with humans in all aspects of science.

ACCELERATING SCIENCE: THE DRIVERS

New technology in sensors, detectors, sequencing, imaging and simulation offers unprecedented opportunities for not only accelerating scientific advances, but also enabling new modes of scientific discovery. New scientific advances in many disciplines are increasingly being driven by our ability to acquire, share, integrate and analyze disparate types of data (Gil et al., 2014), leading to what has been suggested to be a new scientific paradigm, namely data-intensive science (Hey, Tansley, and Tolle, 2009). The resulting challenges in storage, organization, curation, access, sharing, management, processing, analytics, statistics, and visualization s are widely recognized and form the focus of much current research. Modern data analytics techniques that integrate sophisticated probabilistic models, statistical inference, and scalable data structures and algorithms into machine learning systems 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; physical scientists to improve our basic understanding of the physical world, 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. However, despite, and perhaps because of, advances in “big data” technologies for data acquisition, management and analytics, (bottom left of Figure 1), the other largely manual, and labor-intensive aspects the scientific process (the rest of Figure 1) have become the rate limiting steps in scientific progress.

Consider for example, the task of identifying a question for investigation in a domain of inquiry, e.g., the Life Sciences. This is a non-trivial task that requires a good grasp of the current state of knowledge, the expertise and skills needed, the instruments of observation available, the experimental manipulations that are possible, the data analysis and interpretation tools available, etc. Understanding the current state of knowledge requires mastery of the relevant scientific literature which, much like many other kinds of “big data”, is growing at an exponential rate. For example, in 2011, the number of peer-reviewed biomedical research articles appearing in *Pubmed* exceeded 2700 articles per day. The sheer volume and the rate of growth of scientific literature makes it impossible for a scientist to keep up with advances that might have a bearing on the questions being pursued in his or her laboratory. The magnitude of this challenge is further compounded by the fact that many scientific investigations increasingly need to draw on data from a multitude of databases (e.g., Genbank, Protein Data Bank, etc. in the life sciences) and expertise and results from multiple disciplines.

As another example, consider the task of designing an optimal experiment that provides the most valuable information at the lowest cost to help answer a chosen scientific question requires a careful exploration of the space of possible experiments, their relative cost, risk, and feasibility, in the context of all that is known. This challenge is further compounded by the varying degrees of uncertainty associated with the scientific findings.

The components of the scientific process present similar challenges. This underscores the need for much improved cognitive tools for assisting scientists with the rate-limiting steps of the scientific process.

ACCELERATING SCIENCE: FEASIBILITY

In what follows, we argue that computation increasingly serves as:

- A language for science, a role not unlike that played by mathematics over the past many centuries; and
- A powerful formal framework and exploratory apparatus for the conduct of science. These developments together set the stage for developing the cognitive tools needed to accelerate science.

Computing as a language of science

It was nearly a century ago that Rutherford said “All science is either stamp collecting or physics”. Advances in computing, storage, and communication technologies have made it possible to organize, annotate, link, share, and analyze increasingly voluminous, exquisitely diverse data, or in Rutherford’s words, ‘stamp collections’. Recall that 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, and the Newton’s laws that specified how they related to each other, that helped transform the study of the physical universe from “stamp collecting” to “physics,” from a descriptive science into a predictive science. While whether there exist analogs of the simple laws of classical physics for complex biological, cognitive, economic, and social systems might be debatable, that the invention of calculus by Newton and Leibnitz is what made possible the emergence of physics is not.

Mathematics is generally regarded as the language of science. *Algorithms*—precise recipes that describe the relationships between and the processes that operate on the entities that make up the world around us—offer a means for expressing *constructive* mathematics⁴. Algorithms allow us to at least

⁴ Constructive mathematics is distinguished from its traditional counterpart, classical mathematics, by the strict interpretation of the phrase “there exists” as “we can construct” (Bridges and Palmgren, 2013).

approximate *anything* that is describable, including highly non-linear phenomena that cannot be described using equations that have closed-form solutions. There is a growing recognition that processes of interest in biological, social, and cognitive sciences can be viewed as essentially information processes. Arguably, "applied computer science is now playing the role which mathematics did from the seventeenth through the twentieth centuries: providing an orderly, formal framework and exploratory apparatus for other sciences" (Djorgovski, 2005).

This allows us to examine biological, cognitive, and social processes through a *computational lens*, that is, in terms of information processing abstractions (Karp, 2011). Hence, 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 3-dimensional structure of the protein (or more precisely, a set of stable configurations). Examination of natural processes through the computational lens sheds new light on old scientific problems in the respective scientific disciplines. For example, Holland's and Valiant's examinations of biological evolution through the computational lens provide new insights into evolution of complex organisms (Holland, 1975; Valiant, 2009). Roughgarden's work shows how computational complexity sheds new light on the "bounded rationality" of decision-makers (Roughgarden, 2010). Kleinberg's work has provided a new perspective on fundamental questions, e.g., the small world phenomenon (Kleinberg, 2000), in the social sciences.

Examination of a natural process through the computational lens necessarily requires algorithmic or information processing abstractions of the relevant natural entities, relations, and processes. Once such abstractions are created, they become first class computational artifacts in their own right that can be analyzed, shared, and integrated with other related artifacts, contributing to the acceleration of science.

Computing as a formal framework for science

The scientific enterprise (See Figure 1), entails acquiring, organizing, verifying, validating, integrating, analyzing, reasoning with, and communicating descriptions of scientific artifacts, namely, experiments, data, models, hypotheses, theories, and explanations associated with natural or built systems lie at the heart of the scientific enterprise. Hence, computing, the science of information processing, offers a powerful formal framework and exploratory apparatus for the conduct of science (Djorgovski, 2005). It also offers the theoretical and experimental tools for the study of the feasibility, structure, expression, and, when appropriate, automation of (aspects of) the scientific process, the structure and organization of collaborative teams, modeling the evolution of scientific disciplines, and measuring the impact of scientific discoveries.

Accelerating science through automation of aspects the scientific process has been a topic of considerable interest in computer science (Duda et al., 1979; deJong and Rip, 1987; Langley, 1981; Langley et al., 1987; Lindsay et al., 1980; Dzeroski et al., 2007; Shrager and Langley, 1990; Valdez-Perez, 1999; Bradley et al., 2001; Glymour, 2004) as well as cognitive science (Klahr, 2000). Intelligent software agents are already widely used in many aspects of scientific activity. However, this work falls short of accelerating most aspects of science (Waltz and Buchanan, 2009).

Recent advances in robotics for data acquisition, data bases and knowledge bases that capture the relevant background knowledge in specific disciplines, open access to large bodies of scientific literature, technologies for connecting resources and experts, and for constructing and sharing scientific workflows have led to a renewed interest in the topic. For example, King et al. (2009) have demonstrated a robot scientist capable of generating and testing hypotheses, and choosing the experiment to try next, to understand the functional genomics of yeast (*S. cerevisiae*). Schmidt and Lipson (2009) have demonstrated a system that discovers compact equations describe complex nonlinear dynamical

systems, from observations. These demonstrations suggest the possibility of accelerating science by automating some aspects of the scientific process.

ACCELERATING SCIENCE: A RESEARCH AGENDA IN COMPUTER AND INFORMATION SCIENCES

Accelerating science to keep pace with the rate of data acquisition and data processing calls for concerted research efforts that encompass both:

- Development, analysis, integration, sharing, and simulation of algorithmic or information processing abstractions of natural processes, coupled with formal methods and tools for their analyses and simulation;
- Innovations in cognitive tools that augment and extend human intellect and partner with humans in all aspects of science.

In what follows, we elaborate on each of these in turn.

Algorithmic Abstractions for Accelerating Science

The success of computational lens in shedding new light on long-standing questions in biological, cognitive, and social sciences is contributing to their transformation from descriptive sciences into predictive sciences. However, in most disciplines, this transformation is far from complete. In many areas, such abstractions are scarce. In others, the abstractions and the hypotheses that they offer have remained untested, at least in part, due in part to the limitations of our instruments of observation and experimentation and in part due to the cost and complexity of the scientific enterprise. In order for a broad range of sciences and scientists to benefit from the use of computational lens in their respective disciplines, there is an urgent need for developing, sharing, analyzing, and integrating computational abstractions or representations of the key entities, relationships, and processes of interest in the respective scientific disciplines. For example, progress in life sciences has been accelerated substantially with the emergence of gene ontology (Ashburner et al., 2000). Much work remains to be done in a similar vein in other scientific disciplines. Of particular interest are system-level, mechanistic, computational models of biological, cognitive, and social systems that enable the integration of different processes into coherent and rigorous representations that can be analyzed, simulated, integrated, shared, validated against experimental data, and used to guide experimental investigations. Such abstractions, coupled with formal methods for their analysis, can provide rich defined modeling languages with precise syntax and semantics that can be analyzed systematically and efficiently for certain properties of interest. For example, a question of interest to a cancer biologist, e.g. whether the up-regulation of genes A and B and down-regulation of gene C could possibly take a cell from a healthy state to a cancerous state can be translated into a reachability query against a model of a cell where the state of the cell encodes the expression levels of the genes. While there has been some progress in developing such abstractions for molecular and systems biology (Priami, 2009; Bernot et al., 2004; Danos and Laneve, 2004; Fisher and Henzinger, 2007), much work remains to be done, especially in relation to formalisms that allow specification of models that take into account uncertainty and variability, as well as couplings across multiple levels of abstraction, e.g., molecules, cells, tissues, organs, organisms. Similar advances are needed in other scientific disciplines. Of particular interest are formalisms for bridging models not only across levels of abstraction, but also, disciplinary boundaries, to allow studies of complex interactions, e.g., those that couple food, energy, water, environment, and people.

Cognitive Tools for Accelerating Science

In order for science to keep pace with the rate of data acquisition and data processing, there is an urgent need for innovations in cognitive tools (Saloman et al., 1991) for scientists, i.e., computational

tools that leverage and extend human intellect (Engelbart, 1962), and partner with humans on a broader range of tasks involved in scientific discovery (formulating a question, designing, prioritizing and executing experiments designed to answer the question, drawing inferences and evaluating the results, and formulating new questions, in a closed-loop fashion). This calls for deeper understanding formalization, and algorithmic abstractions of, various aspects of the scientific process; development of the computational artifacts (representations, processes, software) that embody such understanding; and the integration of the resulting artifacts into collaborative human-machine systems to advance science (by augmenting, and whenever feasible, replacing individual or collective human efforts). The resulting computer programs would need to close the loop from designing experiments to acquiring and analyzing data to generating and refining hypotheses back to designing new experiments.

Accelerating science calls for programs that can access and ingest information and background knowledge relevant to any scientific question. As search engines and digital libraries return more articles in response to a query than anyone can read, e.g., Google returns about 3.67 million hits for “cancer biology”, there is a need for programs that can read, assess the quality and trustworthiness of, and interpret such information. Machine reading (Etzioni et al., 2006) and information extraction (Niu et al., 2012) are already active areas of research in computer science that have been successfully applied in the life sciences (Hunter and Cohen, 2006; Mallory et al., 2015) and found their way into commercial technology such as IBM Watson (Ferrucci, et al., 2013). Of particular interest are methods for extracting scientific claims from literature and linking them to supporting assumptions, observations or experiments, answering questions, quantifying uncertainty associated with the answers, etc.

Another active area of research is literature-based discovery (Swanson and Smalheiser, 1997; Smalheiser, 2012), which has had some success in finding new relationships between existing knowledge from literature spanning two or more topics (Cameron et al., 2013). Other work on text analytics has led to powerful methods for understanding the evolution of scientific disciplines (Börner et al., 2004; Sinatra et al., 2015), recommending collaborators (Chen et al., 2011), and choosing experiments to accelerate collective discovery (Rzhetsky et al., 2015). However, many challenges remain, e.g., drawing inferences from disparate collections of literature, and increasingly, scientific databases and knowledge bases that contain information of varying degrees of quality and reliability, tracking the evolution of disciplines, identifying major gaps in scientific knowledge, and areas ripe for breakthroughs.

An emerging area of research focuses on data driven approaches to characterizing, and modeling the evolution of scientific disciplines. For example, the results of a recent analysis of the Physics literature (Sinatra et al., 2015) calls into question the conventional narrative of physics as one of paradigm shifts (Kuhn, 1996), divorced from other sciences, and shows that physics has always been in a constant dialog with other disciplines from mathematics to chemistry and even theology, a dialog that is largely driven by the idea that complex phenomena can be understood in terms of a small number of universal laws. Such analyses could allow us to understand the evolution of scientific disciplines, and the impact of a scientific discovery within and beyond the discipline, and identify unexplored areas that are ripe for investigation.

With the exponential growth in scientific literature, often with conflicting scientific arguments, supported by observations of variable quality and analyses made under differing assumptions, there is a dire need for tools for managing conflicting arguments, tracking changes in the validity of the observations and assumptions that they rely on, and support justifiable conclusions. While there is considerable work on computational argumentation systems (Besnard and Hunter, 2008), much work is needed to develop argumentation formalisms and tools that can help accelerate science. Of particular interest are expressive yet computationally tractable languages for representing and reasoning with scientific arguments, and their uncertainty and provenance.

A shift in emphasis from accelerating data collection and data processing to accelerating the entire scientific process calls for representation and modeling languages with precise formal semantics for describing, sharing, and communicating scientific observations (including measurement models) experiments, data, models, theories, conjectures, and hypotheses. The increasing reliance on cognitive tools requires that the all of these be specified in a form that can be processed by computers; and queries against them be translated into precise computational problems.

Even the relatively mundane task of data collection presents many questions including deciding which variables to measure, why, and how i.e., the instrument to use (if one exists) or to design (if need be). Scientific workflows (Gil et al., 2007; Davidson and Freire, 2008) already provide useful ways to describe, manage, share, track data provenance within, and reproduce complex scientific analyses. Scientific systems are already being used for data analyses in the life sciences (Hull et al., 2006). However, there is a need for languages and tools for describing the measurement process, the data models for describing observations using standard ontologies (when they exist), establishing semantics preserving mappings across data models. There is an urgent need for precise languages and tools for describing experiments, methods for quantifying the marginal utility of experiments, determining the scientific as well as economic feasibility of experiments, comparing alternative experiments, and choosing optimal experiments (in a given context). The same holds for hypotheses, conjectures, theories, scientific workflows, and other scientific artifacts.

Machine learning currently offers one of the most cost-effective approaches to constructing predictive models from data (Ghahramani, 2015; Jordan and Mitchell, 2015) across a number of disciplines including biological sciences (Baldi and Brunak, 2001), brain sciences (Pereira et al., 2009), learning sciences (Romero and Ventura, 2010), biomedical and health sciences (Jensen et al., 2012), environmental science (Hampton et al., 2013), and climate science (Faghmous et al., 2014). For example, in biological sciences, machine learning algorithms are routinely used to build predictors of gene structure (McAuliffe et al., 2004), molecular interactions and interfaces (Xue et al., 2015), and to uncover regulatory interactions between genes (Segal et al., 2003). However, such models are often complex hard for scientists to comprehend, and therefore to use to gain mechanistic insights into the underlying phenomena. Consider for example, a support vector machine using a non-linear kernel that predicts whether a target gene of interest is turned on or off based on the previous states of a few hundred other genes. Such a model, its high predictive accuracy, is virtually useless with regard to helping to uncover the underlying genetic regulatory network. There has been some progress in extracting comprehensible knowledge from complex predictive models (Pazzani et al., 1997). A related topic in which there has been considerable interest has to do with methods for incorporating prior knowledge into machine learning (Heckerman et al., 1995; Fung et al., 2002; Cohen, 2014, Faghmous et al., 2014) as well as cognitive modeling (Tenenbaum et al., 2006). However, there remains a significant language gap between model builders and model users. This language gap presents challenges in exploiting prior knowledge to guide model construction, and in interpreting predictive models produced by machine learning in advancing scientific understanding of the underlying domain. For example, in life sciences, directed labeled graph representations of gene regulatory networks (Honavar, 2013) wherein nodes denote genes and directed edges denote regulatory influences, and the + or - labels on the edges denote the excitatory or inhibitory influences are likely to be much more useful in refining our understanding the underlying process, and in suggesting further experiments, than a black box support vector machine with a complex nonlinear kernel that provides the same prediction. Hence, there is an urgent need for a new generation of machine learning algorithms that that can incorporate prior knowledge and constraints from a variety of sources, e.g., from physics, and produce models are expressed in forms that are easy to communicate to disciplinary scientists.

There has been much progress on methods and tools for integrating data from disparate data sources (Lenzerini, 2002; Doan et al., 2012; Haas, 2015); describing data semantics using expressive yet tractable fragments of logic (Berners-Lee et al., 2001; Baader and Nutt, 2003; Calvanese et al., 2007; Horrocks et al., 1999), and more recently, on the more complex problem of sharing knowledge across disparate knowledge bases (Bao et al., 2009; Cuenca Grau et al., 2008; Kutz et al., 2004; Borgida and Serafini, 2003). The resulting semantic web technologies offer powerful tools to support interdisciplinary collaborations within and across scientific disciplines (Hendler, 2003). Yet many challenges remain, especially as they relate to integration of data and knowledge at different levels of abstraction, of differing degrees of uncertainty, trustworthiness, etc.

There has been much recent progress in causal discovery (Spirtes et al., 2000; Pearl, 2009; Illari et al., 2011; Kleinberg, 2012; Berzuiini et al., 2012). There is a growing recognition that answering complex questions increasingly requires synthesizing the findings from data from disparate studies. Conclusions that are obtained in a laboratory setting may not hold exactly a setting that differs in many aspects from that of the laboratory. Often, individual studies, for practical reasons e.g., cost, complexity of the studies, focus on the relationship between a selected set of experimental variables and a specific outcome variable. This means arriving at meaningful answers to questions of interest invariably requires synthesize the findings from multiple such studies, carried out under related, but different experimental settings, under possibly different experimental constraints (e.g., experiments that can be performed on a mouse cannot be carried out on human subjects). While causal discovery from disparate observations and experiments is an active topic of research (e.g., Bareinboim et al., 2013), a great deal of work is needed to characterize the precise conditions under which findings of disparate observational and experimental studies can be synthesized, and to develop cognitive tools for synthesizing such findings when it is appropriate to do so. Also of interest are methods for causal inference across models at multiple levels of abstraction from complex longitudinal data.

While we have effective tools to assist scientists in routine aspects of data management and analytics, barring a few proof-of-concept demonstrations (e.g., King et al., 2009), most of the other steps in the scientific process currently constitute rate limiting steps in scientific progress. These include: Characterizing the current state of knowledge in a discipline and identifying the gaps in the current state of knowledge; Generating and prioritizing questions that are ripe for investigation based on the current scientific priorities and the current state of knowledge; Designing, prioritizing, planning, and executing experiments; Analyzing and interpreting results; Generating and verifying hypotheses; Drawing and justifying conclusions; Validating scientific claims; Replicating studies; Documenting studies; Recording scientific workflows and tracking provenance of data and results; Reviewing and Communicating results; Integrating results into the larger body of knowledge within or across disciplines. Hence, accelerating science requires a rich model of the entire scientific process (See Figure 1) as well as deep knowledge of the scientific area under investigation (Honavar, 2014).

Because science is increasingly a collaborative endeavor, we need: sharable and communicable representations and processes, as well as organizational and social structures and processes, that facilitate collaborative science, including mechanisms for sharing data, experimental protocols, analysis tools, data and knowledge representations, abstractions, and visualizations, tasks, mental models, scientific workflows, mechanisms for decomposing tasks, assigning tasks, integrating results, incentivizing participants, and engaging large numbers of participants with varying levels of expertise and ability in the scientific process through citizen science (Gill and Hirsh, 2012; Bonney et al., 2014).

SUMMARY AND RECOMMENDATIONS

The recent advances in sensing, measurement, storage and communication technologies and the resulting emergence of “big data” offer unprecedented opportunities for not only accelerating scientific advances, but also enabling new modes of discovery. Scientific progress in many disciplines is increasingly driven by advances in our ability to:

- Examine natural phenomena through the computational lens, i.e., using algorithmic or information processing abstractions of the underlying processes;
- Acquire, share, integrate, analyze, and build predictive and causal models from disparate types of data.

However, there is a huge gap between our ability to acquire, store, and process data and our ability to make effective use of the data to advance science. Despite successful automation of routine aspects of data management and analytics, most elements of the scientific process currently constitute rate-limiting steps in the scientific process.

Accelerating science to keep pace with the rate of data acquisition and data processing calls for focused investments in a research program that encompasses both:

- Development, analysis, integration, sharing, and simulation of algorithmic or information processing abstractions of natural processes, coupled with formal methods and tools for their analyses and simulation;
- Innovations in cognitive tools that augment and extend human intellect and partner with humans in all aspects of science. This requires:
 - The formalization, development, analysis, of algorithmic or information processing abstractions of various aspects of the scientific process;
 - The development of computational artifacts (representations, processes, software) that embody such understanding; and
 - The integration of the resulting cognitive tools into collaborative human-machine systems and infrastructure to advance science.

Of particular urgency are investments in:

- Algorithmic abstractions of:
 - The natural entities, relations, and processes of interest in specific scientific disciplines;
 - Formal methods and tools for their analyses and simulation;
 - Formalisms for specification of models that take into account uncertainty, and variability;
 - Couplings across multiple levels of abstraction and spatial and temporal granularity;
- Cognitive tools for:
 - Mapping the current state of knowledge in a discipline and identifying the major gaps;
 - Generating and prioritizing questions that are ripe for investigation based on the current scientific priorities and the gaps in the current state of knowledge;
 - Machine reading, including methods for extracting and organizing descriptions of experimental protocols, scientific claims, supporting assumptions, and validating scientific claims from scientific literature, and increasingly scientific databases and knowledge bases;
 - Literature-based discovery, including methods for drawing inferences and generating hypotheses from existing knowledge in the literature (augmented with discipline-specific

databases and knowledge bases of varying quality when appropriate), and ranking the resulting hypotheses;

- Expressing, reasoning with, updating scientific arguments (along with supporting assumptions, facts, observations), including languages and inference techniques for managing multiple, often conflicting arguments, assessing the plausibility of arguments, their uncertainty and provenance;
 - Observing and experimenting, including languages and formalisms for describing and harmonizing the measurement process and data models, capturing and managing data provenance, describing, quantifying the utility, cost, and feasibility of experiments, comparing alternative experiments, and choosing optimal experiments (in a given context);
 - Navigating the spaces of hypotheses, conjectures, theories, and the supporting observations and experiments;
 - Analyzing and interpreting the results of observations and experiments, including machine learning methods that: explicitly model the measurement process, including its bias, noise, resolution; incorporate constraints e.g., those derived from physics, into data-driven inference; close the gap between model builders and model users by producing models that are expressible in representations familiar to the disciplinary scientists;
 - Synthesizing, in a principled manner, the findings in a target setting from disparate experimental and observational studies (e.g., implications to human health of experiments with mouse models);
 - Documenting, sharing, reviewing, replicating, and communicating entire scientific studies in the form of reproducible and extensible scientific workflows;
 - Communicating results of scientific studies and integrating the results into the larger body of knowledge within or across disciplines;
 - Collaborating, communicating, and forming teams with other scientists with complementary knowledge, skills, expertise, and perspectives on problems of common interest (including problems that span disciplinary boundaries or levels of abstraction);
 - Organizing and participating in citizen science projects, including tools for decomposing tasks, assigning tasks, integrating results, incentivizing participants, and engaging large numbers of participants with varying levels of expertise and ability in the scientific process;
 - Cognitive tools for tracking scientific progress, the evolution of scientific disciplines and scientific impact.
- Multi-disciplinary, interdisciplinary, and trans-disciplinary teams that bring together:
 - Experimental scientists in a discipline, e.g., the biomedical sciences, with information and computer scientists, mathematicians, etc., to develop algorithmic or information processing abstractions to support theoretical and experimental investigations;
 - Organizational and social scientists and cognitive scientists to study such teams, learn how best to organize and incentivize such teams and develop a science of team science;
 - Experimental scientists in one or more disciplines, computer and information scientists and engineers, organizational and social scientists, cognitive scientists, and philosophers of science to design, implement, and study end-to-end systems that flexibly integrate the relevant cognitive tools into complex scientific workflows to solve broad classes of problems in specific domains, e.g., understanding complex interactions between food, energy, water, environment, and populations.

- Interdisciplinary graduate and undergraduate curricula and research based training programs to prepare:
 - A diverse cadre of computer and information scientists and engineers with adequate knowledge of one or more scientific disciplines to design, construct, analyze and apply algorithmic abstractions, cognitive tools, and end-to-end scientific workflows in those disciplines;
 - A new generation of natural, social, and cognitive science researchers and practitioners fluent in the use of algorithmic abstractions and cognitive tools to dramatically accelerate and explore new modes of discovery within and across disciplines.

A research agenda focused on accelerating science can be expected to yield:

- Fundamental advances in multiple areas of computer and information sciences, including, theory of computation, complexity theory, algorithms, formal methods, knowledge representation and inference, information integration, machine reading, software engineering, machine learning, causal inference, multi-objective optimization, argumentation systems, planning, decision making, computational organization theory, robotics, human-computer-robot interaction, among others;
- Cognitive tools that could dramatically accelerate scientific progress, by leveraging and extending the reach of human intellect, and partnering with scientists, including citizen scientists, with a broad range of skills and expertise.

This white paper has sought to articulate a research agenda for developing cognitive tools that can augment human intellect and partner with humans on the scientific process. The resulting new cognitive tools can help realize the transformative potential of big data in many sciences, by dramatically accelerating science. The benefits of accelerating science extend well beyond the scientific community to all of humanity: Precision health regimens that take into account not only one's genetic makeup, but also environment, and lifestyle; Personalized education that optimizes curriculum, pedagogy, etc. to optimize the learning outcomes for each individual; Precision agriculture that optimizes everything from the choice of crops to water and fertilizer use to optimize yield and impact on the environment.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. (1136993). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The article has benefited from discussions with and feedback from the broader scientific community, and current and former members of the Computing Community Consortium Council, especially Greg Hager, Cynthia Dwork, Liz Bradley, Klara Nahrstedt, Elizabeth Mynatt, Ben Zorn, Susan Davidson, Susan Graham.

REFERENCES

1. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., and J.T. Eppig, J.T. (2000). Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25:25–29
2. Baader, F. and Nutt, W. (2003). Basic description logics. In *Description logic handbook* pp. 43-95. Springer.
3. Baldi, P. and Brunak, S., 2001. *Bioinformatics: the machine learning approach*. MIT press.
4. Bao, J., Voutsadakis, G., Slutzki, G. and Honavar, V. (2009). Package-based description logics. In *Modular Ontologies* (pp. 349-371). Springer Berlin Heidelberg.

5. Bareinboim, E., Lee, S., Honavar, V. and Pearl, J. (2013). Transportability from multiple environments with limited experiments. In *Advances in Neural Information Processing Systems*: pp. 136-144.
6. Berners-Lee, T., Hendler, J. and Lassila, O. (2001). The semantic web. *Scientific American*, 284:28-37.
7. Bernot G, Comet JP, Richard A, Guespin J. (2004). Application of formal methods to biological regulatory networks: extending Thomas' asynchronous logical approach with temporal logic. *Journal of theoretical biology*. 229(3): 339-47.
8. Berzuini, C., Dawid, P. and Bernardinell, L. eds. (2012). *Causality: Statistical perspectives and applications*. John Wiley & Sons.
9. Besnard, P., and Hunter, A. (2008) *Elements of argumentation*. Cambridge: MIT press.
10. Bonney, R., Shirk, J.L., Phillips, T.B., Wiggins, A., Ballard, H.L., Miller-Rushing, A.J. and Parrish, J.K., (2014). Next steps for citizen science. *Science*, 343(6178), pp.1436-1437.
11. Borgida, A. and Serafini, L. (2003). Distributed description logics: Assimilating information from peer sources. *Journal on data semantics* 1: 153-184.
12. Börner, K., Maru, J.T. and Goldstone, R.L., 2004. The simultaneous evolution of author and paper networks. *Proceedings of the National Academy of Sciences*, 101 (suppl 1): 5266-5273.
13. Bradley E, Easley M, Stolle R. (2001) Reasoning about nonlinear system identification. *Artificial Intelligence*. 2001 Dec 31;133(1):139-88.
14. Bridges, D. and Palmgren, E. (2013), *Constructive Mathematics*. In: *The Stanford Encyclopedia of Philosophy* Edward N. Zalta (ed.), <http://plato.stanford.edu/archives/win2013/entries/mathematics-constructive/>.
15. Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M. and Rosati, R. (2007). Tractable reasoning and efficient query answering in description logics: The DL-Lite family. *Journal of Automated reasoning*, 39:385-429.
16. Cameron, D., Bodenreider, O., Yalamanchili, H., Danh, T., Vallabhaneni, S., Thirunarayan, K., Sheth, A.P. and Rindflesch, T.C. (2013). A graph-based recovery and decomposition of Swanson's hypothesis using semantic predications. *Journal of biomedical informatics*, 46:238-251.
17. Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka Jr, E.R. and Mitchell, T.M. (2010). Toward an Architecture for Never-Ending Language Learning. In *AAAI* pp. 1306-1313.
18. Chalmers, A.F. (1999). *What is this thing called Science?* University of Queensland Press.
19. Chen, H.H., Gou, L., Zhang, X. and Giles, C.L., 2011, June. Collabseer: A Search Engine for collaboration discovery. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries* pp. 231-240. ACM.
20. Cohen, W.W. (2014) Compiling prior knowledge into an explicit bias. In *Proceedings of the Ninth International Conference on Machine Learning* pp. 102-110.
21. Cuenca Grau, B., Horrocks, I., Kazakov, Y. and Sattler, U. (2008). Modular reuse of ontologies: Theory and practice. *Journal of Artificial Intelligence Research*, pp. 273-318.
22. Danos V, Laneve C. (2004). Formal molecular biology. *Theoretical Computer Science*. 325(1): 69-110.
23. Davidson, S.B. and Freire, J. (2008). Provenance and scientific workflows: challenges and opportunities. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pp. 1345-1350. ACM.
24. de Jong, H. and Rip, A. (1997). The Computer Revolution in Science: Steps Towards the Realization of Computer-Supported Discovery Environments *Artificial Intelligence* 91: 225-256.
25. Djorgovski (2005). V. In: *Proc. of CAMPO5, "Computer Architectures for Machine Perception"*, Di Gesu & D. Tegolo (Ed). IEEE Press. pp. 125-132.
26. Doan, A., Halevy, A. and Ives, Z. (2012). *Principles of data integration*. Elsevier.

27. Duda, R.O., Gaschnig, J., Hart, P.E. (1979) in: *Expert Systems in the Microelectronic Age*, D. Michie, Ed. Edinburgh Univ.Press. pp. 153–167.
28. Dzeroski, S., Langley, P., and Todorovski, L. (2007). *Computational Discovery of Scientific Knowledge*. In: Dzeroski, S., Todorovski, L. (Eds.) (2007), *Computational discovery of communicable scientific knowledge*. Berlin: Springer.
29. Engelbart, D.C., 2001. *Augmenting human intellect: a conceptual framework (1962)*. PACKER, Randall and JORDAN, Ken. *Multimedia. From Wagner to Virtual Reality*. New York: WW Norton & Company, pp.64-90.
30. Faghmous, J.H., Banerjee, A., Shekhar, S., Steinbach, M., Kumar, V., Ganguly, A.R. and Samatova, N. (2014). Theory-guided data science for climate change. *Computer*, 11:74-78.
31. Ferrucci, David A., Anthony Levas, Sugato Bagchi, David Gondek, and Erik T. Mueller. *Watson: beyond jeopardy!* *Artif. Intell.* 199 (2013): 93-105.
32. Fisher J, Henzinger TA. (2007). Executable cell biology. *Nature biotechnology*. 25(11): 1239-49.
33. Fung, G.M., Mangasarian, O.L. and Shavlik, J.W. (2002). Knowledge-based support vector machine classifiers. In *Advances in neural information processing systems* pp. 521-528.
34. Etzioni, Oren, Michele Banko, and Michael J. Cafarella. *Machine Reading*. In *AAAI*, pp. 1517-1519. 2006.
35. Gil, Y., Deelman, E., Ellisman, M., Fahringer, T., Fox, G., Gannon, D., Goble, C., Livny, M., Moreau, L. and Myers, J. (2007). Examining the challenges of scientific workflows. *IEEE Computer*, 40(12): 26-34.
36. Gil, Y. and Hirsh, Y. (2012). *Discovery Informatics: AI Opportunities in Scientific Discovery*. *AAAI Technical Report FS-12-03*.
37. Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, 521:452-459.
38. Glymour, C. (2004). *The automation of Discovery*. *Daedalus Winter 2004*, pp. 69-77.
39. Haas, L.M. (2015) *The Power Behind the Throne: Information Integration in the Age of Data-Driven Discovery*. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. pp. 661-661). ACM.
40. Hacking, I. (1983). *Representing and Intervening. Introductory Topics in the Philosophy of Science*. Cambridge University Press.
41. Hampton, S.E., Strasser, C.A., Tewksbury, J.J., Gram, W.K., Budden, A.E., Batcheller, A.L., Duke, C.S. and Porter, J.H. (2013). Big data and the future of ecology. *Frontiers in Ecology and the Environment*, 11:156-162.
42. Heckerman, D., Geiger, D. and Chickering, D.M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine learning*, 20:197-243.
43. Hendler, J. (2003). Science and the Semantic Web. *Science* 299:520-521.
44. Gil, Y., Greaves, M., Hendler, J., and Hirsh, H. (2014). Amplify scientific discovery with artificial intelligence. *Science* 346:171-172.
45. Hey, Tansley, S. and Tolle, K.M. (Ed) (2009). *The fourth paradigm: data-intensive scientific discovery*. Redmond, WA: Microsoft Research
46. Holland, J.H. (1975). *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. U Michigan Press.
47. Honavar, V. (2013). *From Data Analytics to Discovery Informatics*. In: *Data Science: Unlocking the Power of Big Data*. National Institutes of Health Videocast. <http://videocast.nih.gov/summary.asp?bhjs=0&File=17798>.

48. Honavar, V. (2014). The promise and potential of big data: A case for discovery informatics. *Review of Policy Research*, 31(4), pp.326-330.
49. Horrocks, I., Sattler, U. and Tobies, S. (1999), September. Practical reasoning for expressive description logics. In *Logic for Programming and Automated Reasoning*. pp. 161-180. Springer Berlin Heidelberg.
50. Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Pocock, M.R., Li, P. and Oinn, T., 2006. Taverna: a tool for building and running workflows of services. *Nucleic acids research*, 34 (suppl 2), pp. W729-W732.
51. Hunter, L. and Cohen, K.B. (2006). Biomedical language processing: what's beyond PubMed? *Molecular cell*, 21(5): 589-594.
52. Illari, P., Russo, F. and Williamson, J. eds. (2011). *Causality in the Sciences*. Oxford University Press.
53. Jensen, P.B., Jensen, L.J. and Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13:395-405.
54. Jordan, M.I. and Mitchell, T.M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349:255-260.
55. Karp RM. (2011). Understanding science through the computational lens. *Journal of Computer Science and Technology*, 26(4): 569-77.
56. Klahr, D. (2000). *Exploring Science: The Cognition and Development of Discovery Processes*, Cambridge, MA: MIT Press.
57. Kleinberg J. (2000). The small-world phenomenon: An algorithmic perspective. In: *Proceedings of the thirty-second annual ACM symposium on Theory of computing*. pp. 163-170. ACM.
58. Kleinberg, S. (2012). *Causality, probability, and time*. Cambridge University Press.
59. Kuhn, T. S. (1996). *The Structure of Scientific Revolutions* (3rd ed.). University of Chicago Press.
60. Kutz, O., Lutz, C., Wolter, F. and Zakharyashev, M. (2004). E-connections of abstract description systems. *Artificial intelligence*, 156: 1-73.
61. Langley, P. (1981) Data - driven discovery of physical laws. *Cognitive Science*, 5(1), pp.31-54.
62. Langley, P., Simon, H.A., Bradshaw, G.L., and Zytkow, J.M. (1987). *Scientific discovery: Computational explorations of the creative processes*. Cambridge, MA: MIT Press.
63. Lenzerini, M. (2002). Data integration: A theoretical perspective. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* pp. 233-246. ACM.
64. Lindsay, R.K., Buchanan, B.G., Feigenbaum, E.A., and Lederberg, J. (1980). *Applications of Artificial Intelligence for Organic Chemistry: The DENDRAL Project*, McGraw-Hill.
65. Mallory EK, Zhang C, Ré C, Altman RB. Large-scale extraction of gene interactions from full text literature using DeepDive. *Bioinformatics*. 2015 btv476.
66. McAuliffe, J.D., Pachter, L. and Jordan, M.I. (2004). Multiple-sequence functional annotation and the generalized hidden Markov phylogeny. *Bioinformatics*, 20:1850-1860.
67. Niu, F, Zhang C, Ré C, Shavlik J.W. (2012). DeepDive: Web-scale Knowledge-base Construction using Statistical Learning and Inference. *VLDS 12*:25-8.
68. Pazzani, M.J., Mani, S. and Shankle, W.R., 1997. Beyond Concise and Colorful: Learning Intelligible Rules. In *KDD* (Vol. 97, pp. 235-238).
69. Pearl, J. (2009). *Causality*. Cambridge university press.
70. Pereira, F., Mitchell, T. and Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage*, 45(1): S199-S209.
71. Priami C. (2009). Algorithmic systems biology. *Communications of the ACM*. 52(5): 80-8.
72. Romero, C. and Ventura, S. (2010). Educational data mining: a review of the state of the art. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 40:601-618.

73. Rosenberg, A. (2000). *Philosophy of Science*. Routledge Press.
74. Roughgarden T. (2010). Algorithmic game theory. *Communications of the ACM*. 53(7):78-86.
75. Rzhetsky, A., Foster, J.G., Foster, I.T. and Evans, J.A. (2015). Choosing experiments to accelerate collective discovery. *Proceedings of the National Academy of Sciences*, 112(47):14569-14574.
76. Salomon, G., Perkins, D.N. and Globerson, T. (1991). Partners in cognition: Extending human intelligence with intelligent technologies. *Educational researcher*, 20(3), pp.2-9.
77. Schmidt, M. and Lipson, H. (2009). Distilling free-form natural laws from experimental data. *Science*, 324(5923), pp.81-85.
78. Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D. and Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature genetics*, 34:166-176.
79. Shrager, J., & Langley, P. (Eds.) (1990). *Computational models of scientific discovery and theory formation*. San Mateo, CA: Morgan Kaufmann.
80. Sinatra R, Deville P, Szell M, Wang D, Barabási A.L. (2015). A century of physics. *Nature Physics*. 11(10):791-6.
81. Smalheiser, N.R. (2012), Literature-based discovery: Beyond the ABCs. *J. Am. Soc. Inf. Sci.*, 63: 218–224. doi: 10.1002/asi.21599
82. Spirtes, P., Glymour, C.N. and Scheines, R. (2000). *Causation, prediction, and search*. MIT press.
83. Swanson DR, Smalheiser NR. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence* 1997; 91: 183-203.
84. Tenenbaum, J.B., Griffiths, T.L. and Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in cognitive sciences*, 10:309-318.
85. Valdez-Perez, R.E. (1999). Principles of Human-Computer Collaboration for Knowledge Discovery in Science *Artificial Intelligence* 107:335-346.
86. Valiant, L.G. (2009). Evolvability. *Journal of the ACM (JACM)*, 56(1), DOI = 10.1145/1462153.1462156 <http://doi.acm.org/10.1145/1462153.1462156>
87. Waltz, D. and Buchanan, B.G., 2009. Automating science. *Science*, 324(5923), pp.43-44.
88. Xue, L.C., Dobbs, D., Bonvin, A.M. and Honavar, V. (2015). Computational prediction of protein interfaces: A review of data driven methods. *FEBS letters*, 589:3516-3526.
89. Zytkow, J.M., Zhu, J. and Hussam, A. (1990) Automated Discovery in a Chemistry Laboratory. In *AAAI* pp. 889-894.

This material is based upon work supported by the National Science Foundation under Grant No. (1136993). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.