# Advanced Cyberinfrastructure for Accelerating Science[1]

A whitepaper submitted to the NSF Workshop on Developing a Roadmap towards the Next Generation of Smart Cyberinfrastructure

Vasant G. Honavar[2]
Artificial Intelligence Research Laboratory
Center for Big Data Analytics and Discovery Informatics
Institute for Computational and Data Sciences
Pennsylvania State University

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 laws of planetary motion. Later Isaac Newton used these laws and other data to derive his unified laws of motion and laid 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 scientific process that they engaged in is summarized in Figure 1.

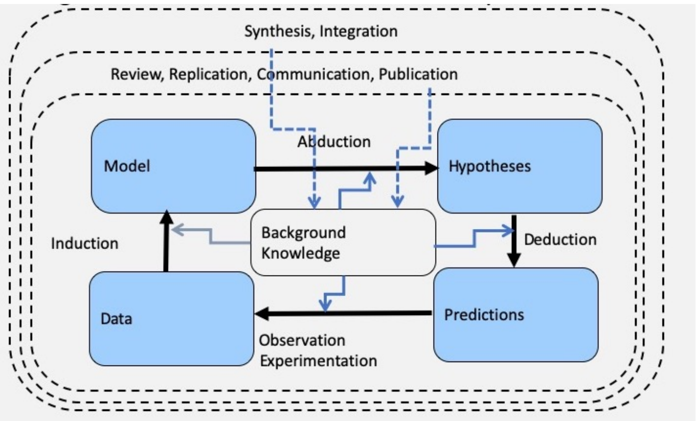

Figure 1: Major Components of the Scientific Process

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

[1] Some of the material in this white paper has been adapted from: (i) Honavar V., Hill M., & Yelick K. (2016). *Accelerating Science: A Computing Research Agenda*. A white paper prepared for the Computing Community Consortium committee of the Computing Research Association. arXiv:1604.02006 and (ii) Honavar V., Yelick K., Nahrstedt K., Rushmeier H., Rexford J., Hill M., Bradley E., & Mynatt E. (2017). Advanced Cyberinfrastructure for Science, Engineering, and Public Policy. A white paper prepared for the Computing Community Consortium committee of the Computing Research Association. arXiv:1707.00599

[2] Professor and Edward Frymoyer Chair of Information Sciences and Technology; Professor, Computer Science, Bioinformatics and Genomics, Informatics, and Neuroscience Graduate Programs; Professor, Data Sciences Undergraduate Program; Director, Artificial Intelligence Research Laboratory; Director, Center for Big Data Analytics and Discovery Informatics; Associate Director, Institute for Computational and Data Sciences; Co-PI, Northeast Big Data Innovation Hub; Co-PI, Virtual Data Collaboratory; Steering Committee Member, Eastern Regional Network

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 information (models, hypotheses, theories, explanations) about natural and built systems lie at the heart of the scientific enterprise. The past centuries have witnessed major scientific breakthroughs as a result of advances in instruments of observation, formalisms for describing the laws of nature, improved tools for calculation, and infusion of concepts, tools, and scientific practices across disciplines.

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, most other aspects of the scientific process (understanding the current state of knowledge, formulating questions, designing studies, assembling and managing research teams, identifying designing, prioritizing, optimizing and executing experiments, organizing and integrating data, knowledge, and assumptions to draw inferences and interpret and explain results) constitute an even greater bottleneck than ever.**

In what follows, I will argue that: **Accelerating science calls for foundational advances in Artificial Intelligence and the translation of the resulting advances into cognitive tools that amplify, augment, and extend human intellect and abilities through advanced cyberinfrastructure for science.**

**Algorithmic Abstractions of Scientific Domains.** Today, algorithmic abstractions increasingly play, across many sciences, the role played by calculus or more generally, mathematics, in the emergence of physics. For example, in biology, 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). In cognitive science, we will have a theory of learning from experience, when we have an algorithm that learns from observations and experiments. Algorithmic abstractions of the relevant natural entities, relations, and processes in a scientific domain (e.g., biology) allow us to examine the domain through the computational lens and formulate and answer scientific questions in the domain in algorithmic terms. Once created, the algorithmic abstractions 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. **The creation of sufficiently expressive, yet practically useful algorithmic abstractions of scientific domains calls for major advances in AI, including in particular, knowledge representation, knowledge elicitation, and machine learning. Cyberinfrastructure for science must provide the necessary tools and infrastructure for the collaborative creation, sharing, and use of algorithmic abstractions of scientific domains.**

**Algorithmic Abstractions of the Scientific Process:** The scientific enterprise (See Figure 1), entails acquiring, organizing, verifying, validating, integrating, analyzing, reasoning with, and communicating information bearing 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, which offers a powerful medium for digital representation and manipulation of information artifacts offers a powerful formal framework and exploratory apparatus for science. 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. **The creation and realization of algorithmic abstractions of the scientific process calls for foundational advances across multiple areas of AI, including knowledge representation, planning, optimization, search, multi-agent communication and coordination, natural language processing, information extraction, machine learning, among others. Cyberinfrastructure for science must provide generalizable, modular, extensible, interoperable infrastructure and tools for accelerating science by (i) whenever feasible, automating aspects of science (well beyond building predictive models from data using machine learning, and compute and data intensive simulations).**

**Cognitive tools for Amplifying, Augmenting, and Extending Human Intellect and Abilities:** Accelerating science requires effective computational 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; 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, and 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 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 modeling the measurement process, its bias, noise, resolution; incorporating constraints e.g., those derived from physics, into data- driven inference; closing 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, e.g., causal relationships from disparate experimental and observational studies. **The development of cognitive tools for scientists requires foundational advances in AI. Realizing the promise and potential of such cognitive tools to accelerate science requires advanced cyberinfrastructure for implementing, validating, deploying, and operating the tools.**

**Advanced Cyberinfrastructure for Collaborative Science:** Because major activities in science increasingly require on collaboration across disciplinary as well as organizational boundaries, there is a need for data and computational infrastructure to support: Organizing and participating in team 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 process; Collaborating, communicating, and forming teams with partners with complementary knowledge, skills, expertise, and perspectives on problems of common interest (including problems that span disciplinary boundaries or levels of abstraction, and call for collaboration across government, industry and academia); Creating and sharing of human understandable and computable representations of the relevant artifacts, including data, experiments, hypotheses, conjectures, models, theories, workflows, etc. across organizational and disciplinary boundaries; Documenting, sharing, reviewing, replicating, and communicating entire studies in the form of reproducible and extensible workflows (with provision for capturing data provenance); Automating the discovery, adaptation, and when needed, assembly of complex analytic workflows from available components; Communicating results of studies or investigations and integrating the results into the larger body of knowledge within or across disciplines or communities of practice; Tracking scientific progress, the evolution of scientific disciplines, and impact on science, engineering, or public policy. **Amplifying, augmenting, and extending human intellect and abilities to accelerate science, calls for fundamental advances in AI, especially, human-machine, machine-machine, and machine mediated human-human collaboration; and advanced cyberinfrastructure collaborative science across disciplinary and institutional boundaries.**

**Transparent, trustworthy, accountable cyberinfrastructure for science:** As scientific advances rely on data (including sensitive data), infrastructure, and tools beyond those that any individual scientist can fully comprehend or manage, there is a need for: Computable data access and usage agreements that can be enforced within a secure cyberinfrastructure; Audit mechanisms that can be used to verify compliance with the applicable data access and use policies; Repositories of data and their usage agreements that can be adapted and reused in a variety of settings; Agile and secure computing and network services and protocols that can accommodate different types and vintages of instruments; Access privileges that are responsive to the changing needs and roles of individuals; Distributed data management systems or virtual federated collaboratories that enable seamless sharing of data, computational resources, analysis tools, and results across disciplinary and organizational boundaries while ensuring compliance with applicable security as well as data access and use policies; Sustainable model for data and long-term preservation of both data and the software needed to make use of it; Data and software provenance and other mechanisms for ensuring transparency and reproducibility of data analysis, modeling, etc., detecting, and correcting for implicit or explicit biases or errors in the data as well as the algorithms. **Accelerating science calls for advanced cyberinfrastructure, policy frameworks, and tools for ensuring the transparency, trustworthiness, and accountability of advanced cyberinfrastructure for science.**

**Education and Training:** Realizing the promise and potential of advances in AI and cyberinfrastructure to accelerate science calls for: a diverse cadre of scientists who combine deep expertise in a scientific domain that have the knowledge and skills to develop and utilize algorithmic abstractions within their scientific domain; interdisciplinary teams of scientists and engineers to design, implement, and study end-to-end systems that flexibly integrate the relevant cognitive tools into complex workflows to solve broad classes of problems in specific domains; organizational, social, behavioral and cognitive scientists to study cyberinfrastructure enabled team science and discover, and translate to practice how best to organize and incentivize such teams to optimize their effectiveness; and organizational changes and funding models that catalyze the acceleration of science through advances in AI and cyberinfrastructure