Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 16:08:28
REFERENCE NO: 279

Author Names & Affiliations

- Vasant Honavar - Pennsylvania State University
- Katherine Yelick - Lawrence Berkeley National Laboratory
- Klara Nahrstedt - University of Illinois Urbana Champaign
- Jennifer Rexford - Princeton University
- Mark Hill - University of Wisconsin
- Elizabeth Bradley - University of Colorado
- Elizabeth Mynatt - Georgia Institute of Technology

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

Computer Science, Material Science, Life Sciences, Health Sciences, Machine Learning, Scientific Computation, High Performance Computing, Artificial Intelligence, Computer Networks, Computer Security, Data Science, Bioinformatics, Health Informatics

Title of Submission

Advanced Cyberinfrastructure for Accelerating Science

Abstract (maximum ~200 words).

Scientific progress in many disciplines is increasingly enabled by our ability to examine natural phenomena through the computational lens, i.e., using computational abstractions of the underlying processes; and our ability to acquire, share, integrate, and analyze disparate types of data. These advances would not be possible without the advanced data and computational infrastructure and tools for data integration, analysis, modeling, and simulation. However, despite, and perhaps because of, advances in “big data” technologies for data acquisition, management and analytics, the other largely manual, and labor-intensive aspects the scientific process, e.g., formulating questions, designing studies, organizing, curating, and integrating data, drawing inferences and interpreting results, have become the rate-limiting steps in scientific progress. Accelerating science requires support for computational abstractions of scientific domains, coupled with methods and tools for their analysis, synthesis, simulation, visualization, sharing, and integration; cognitive tools that leverage and extend the reach of human intellect, and partner with humans on all aspects of science; representations, processes, protocols, workflows that embody computational abstractions of the scientific process; and because science is increasingly a collaborative endeavor, support for organizational and support for team science that transcends disciplinary boundaries.

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

The emergence of “big data” offers unprecedented opportunities for not only accelerating scientific advances but also enabling new modes of discovery. For example, advances in sequencing, imaging, and online text , coupled with domain-specific computational abstractions, and new methods and tools for data integration, analysis, and modeling, are enabling biologists to gain insights into how living systems adapt and thrive; neuroscientists to uncover how brains adapt and learn; health scientists to personalize treatments and interventions to optimize health outcomes; economists to understand markets; education researchers to personalize curricula and pedagogy to optimize learning outcomes; social scientists to study why organizations, societies, and cultures succeed or fail.

These advances would not be possible without the advanced data and computational infrastructure and tools for data integration, analysis, modeling, and simulation. However, despite, and perhaps because of, advances in “big data” acquisition, management and analytics, the other largely manual, and labor-intensive aspects the scientific process, e.g., formulating questions, designing, generating hypotheses, prioritizing and executing studies, organizing, curating, and integrating data, drawing inferences, interpreting results, and evaluating models, 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. 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 one’s 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.

Scientific progress in many disciplines is increasingly enabled by our ability to examine natural phenomena through the computational lens, i.e., using computational abstractions of the underlying processes; and our ability to acquire, share, integrate, and analyze disparate types of data. These advances would not be possible without the advanced data and computational infrastructure and tools for data integration, analysis, modeling, and simulation. However, despite, and perhaps because of, advances in “big data” technologies for data acquisition, management and analytics, the other largely manual, and labor-intensive aspects the scientific process, e.g., formulating questions, designing studies, organizing, curating, and integrating data, drawing inferences and interpreting results, have become the rate-limiting steps in scientific progress. Accelerating science requires support for computational abstractions of scientific domains, coupled with computational methods and tools for their analysis, synthesis, simulation, visualization, sharing, and integration; cognitive tools that leverage and extend the reach of human intellect, and partner with humans on all aspects of science; agile and trustworthy data cyber-infrastructures that connect, manage diverse instruments; representations, processes, protocols, workflows that embody computational abstractions of the scientific process; and because science is increasingly a collaborative endeavor, support for organizational and social structures and processes for team science that transcends disciplinary and institutional boundaries.

As another example, consider the task of designing a new material with some desired properties. This requires a careful exploration of the space of possible material compositions and manufacturing processes that could yield the material with the desired properties, their relative cost, risk, and feasibility, in the context of all that is known in the relevant areas of science and engineering. However, the absence of shared data storage and management services and agreed upon shared vocabularies, metadata standards and ontologies presents a significant hurdle to data curation, data discovery, and data exchange with other scientists, as well as efforts to combine data from simulations, experiments, and observations.

The preceding examples are intended to be illustrative, and by no means exhaustive.

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

We see the need for advances in cyberinfrastructure in several broad thematic areas:

Computational abstractions of scientific domains and formal methods and tools for their analysis, integration, visualization, simulation, sharing, and discovery. A key prerequisite for realizing the full potential of data and computation to accelerate science and enable new modes of discovery is the ability to view scientific domains through a computational lens, that is, in terms of computational abstractions. Of particular interest are system-level, mechanistic, computational models of physical, 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. These models must not only cross levels of abstraction, but also, disciplinary boundaries, to allow studies of complex interactions, e.g., those that couple food, energy, water, environment, and people.

Hardware infrastructure: High end computational capabilities for modeling, simulation, data analysis and inference. The need for such systems to support physical modeling and simulation problems continues to grow, due to the increasing complexity of scientific inquiry, addressing multiphysics problems and increasing dynamic range in space and time. This will drive the need for access to computational systems of the highest scales in order to be scientifically competitive. At the same time, there is additional demand for computational growth in high throughput computations, e.g., millions of materials modeling simulations to screen for materials with particular properties needed for batteries, electronics, and other applications. These simulations require systems similar to those at the largest scales, including tightly coupled parallel architectures with high performance floating point appropriately balanced with memory bandwidth, high speed interconnects (although at a smaller scale), and software to co-schedule parallel jobs across nodes. The growing size and complexity of data from simulations, experiments, and embedded sensors will also drive the need for computational infrastructure. With deep learning approaches alone, access to large amounts of computing to train these networks is as important as access to large data sets. Architectural choices for machine learning, graph analytics or other data analysis problems may be different than simulation, e.g., leveraging lower precision floating point or requiring low latency networks for walking over graphs, although today similar processor architectures (including GPUs) are being used. An important feature of the computational lens model of science is that it involves integration of data, models and prediction, suggesting that systems will need to handle diverse workloads to efficiently support the scientific process.
Data intensive problems will require increased networking with unique demands for tools and systems to serve high-speed flows from large experiments, as well as on-demand computing that are not suited to queue-based scheduling. Complex workflows may also benefit from containerized software, and storage models will need to evolve to address provenance, accessibility and sustained availability.

Cognitive tools for scientists: The next generation cyberinfrastructure for science needs to provide a broad range of cognitive tools for scientists, i.e., computational tools that leverage and extend human intellect, and partner with humans on a broader range of tasks that make up a scientific workflow (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). That is, the cyberinfrastructure needs to support computational abstractions of various aspects of the scientific process; development of the computational artifacts (representations, processes, software) that embody such abstractions; 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, curating, and analyzing data to generating and refining hypotheses back to designing new experiments. Particularly needed are 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, 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 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, e.g., causal relationships from disparate experimental and observational studies (e.g., implications to human health of experiments with mouse models). Because science is increasingly a collaborative endeavor, we need advances in cyberinfrastructure to support: The creation and sharing of human understandable and computable representations of scientific artifacts, including data, experiments, hypotheses, conjectures, models, theories, workflows, etc. across organizational and disciplinary boundaries; Documenting, sharing, reviewing, replicating, and communicating entire scientific studies in the form of reproducible and

extensible scientific 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 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 team 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; Tracking scientific progress, the evolution of scientific disciplines, and scientific impact.

Question 3 Other considerations (maximum ~1200 words, optional): Any other relevant aspects, such as organization, process, learning and workforce development, access, and sustainability, that need to be addressed; or any other issues that NSF should consider.

Trustworthy Data Cyberinfrastructure: Because science increasingly relies on data that are subject to restrictions on access and use, we need: Computable data access and use agreements that can be enforced by the scientific workflow within a secure cyberinfrastructure; Audit mechanisms that can be used to verify compliance with the applicable data access and use agreements; Repositories of data use agreements that can be adapted and reused in a variety of settings;
Agile and secure computing and network services and protocols that can adjust and protect different types and ages of scientific instruments: Many scientific instruments, (e.g., a microscope which is purchased to last a decade or longer), and their software are purchased and upgraded at very different schedule than computing and networking software (operating system, security, network services, which are often upgraded every few months); Time-evolvable access privileges to data that follow the scientific process from creating data during the experiment to publishing data in public repositories (At the beginning of the scientific process, scientists are very protective of their data; However, after publishing the results, scientists are open to publish and share data. Hence, different access controls need to be ensured); Distributed data management systems that enable integrative analyses of data from different disciplines.

Because addressing the scientific challenges that address societal needs requires the deep integration of knowledge, techniques, and expertise that transcend disciplinary boundaries, cyberinfrastructure in support of 21st century science needs to support multi-disciplinary, interdisciplinary, and transdisciplinary 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.

Training the 21st century scientists who can both leverage and contribute to advances in cyberinfrastructure requires 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.

Ensuring that the cyberinfrastructure investments lead to advances in the state-of-the-art in computational and data infrastructure for science requires support for: Operational infrastructure based on the best available computing and information technology; Integrated data cyber-infrastructure that allows the sharing of data and metadata from simulations and experiments on scientific instruments; Experimental infrastructure to explore novel data, computing, and collaborative technologies and platforms, including the basic computer science and engineering advances needed to meet the needs of 21st century science;
Datasets and tools for training and education that include data from successful as well as failed studies, including experiments and computational analyses.

Consent Statement