

Reference ID: 11226901575_Towns

Reference ID: 11226901575_Towns

Submission Date and Time: 12/16/2019 3:46:16 PM

This contribution was submitted to the National Science Foundation in response to a Request for Information, <https://www.nsf.gov/pubs/2020/nsf20015/nsf20015.jsp>. Consideration of this contribution in NSF's planning process and any NSF-provided public accessibility of this document does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government. The content of this submission is protected by the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>).

Consent Statement: "I hereby agree to give the National Science Foundation (NSF) the right to use this information for the purposes stated above and to display it on a publicly available website, consistent with the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>)."

Consent answer: I consent to NSF's use and display of the submitted information.

Author Names & Affiliations

Submitting author: John Towns - University of Illinois at Champaign-Urbana

Additional authors: Kelly Gaither, University of Texas at Austin; Linda Akli, Southeastern Universities Research Association; Bob Sikovitz, University of California San Diego; Phil Blood, Pittsburgh Supercomputing Center; Rich Knepper, Cornell University; Victor Hazlewood, U

Contact Email Address (for NSF use only): (hidden)

Research domain(s), discipline(s)/sub-discipline(s)

astronomy; astrophysics; earth sciences; climate modeling; materials science; environmental sciences; artificial intelligence; bioinformatics; civil engineering; cosmology; cybersecurity; electrical engineering; humanities; multimedia analytics; particle physics; public health

Title of Response

Observation from XSEDE for Infrastructure to Support Data-intensive and Data-Driven Science

Abstract

As a significant socio-technical platform that integrates and coordinates advanced digital services within the national ecosystem to support contemporary science, the NSF-funded XSEDE (the Extreme Science and Engineering Discovery Environment) appreciates the opportunity to respond to this RFI. XSEDE believes that providing resources, support, and other CI capabilities for scaling up data-intensive research is a complex challenge that we can help solve. As an organization crossing many research domains, XSEDE provides an interesting perspective in seeing the challenges faced by many researchers in many domains--in particular, identifying common challenges that we articulate in the responses to the questions.

Question 1 (maximum 400 words) – Data-Intensive Research Question(s) and Challenge(s). Describe current or emerging data-intensive/data-driven S&E research challenge(s), providing context in terms of recent research activities and standing questions in the field. NSF is particularly interested in cross-disciplinary challenges that will drive requirements for cross-disciplinary and disciplinary-agnostic data-related CI.

In 2019, about one third of XSEDE users worked on data-centric projects in artificial intelligence, bioinformatics, civil engineering, cosmology, cybersecurity, electrical engineering, healthcare, humanities, materials science, multimedia analytics, particle physics, and public health. Data formats, analysis methods, and software workflows differ widely, as does the skill level of their users. Many of these users are new to data science, as well as to advanced computing, so they require considerable training and support in order to be successful in their projects. A sampling of those that have data-intensive need either or in the near future include:

- Large Hadron Collider: As physicists look at luminosity in the future, the data volumes to be stored and shared will be huge. Beyond the challenges in simply managing the data, one of the biggest challenges is in processing/analyzing the data. Collision events are busy with a non-linear dependence of time to reconstruct events. This will require investigating novel architectures (CPUs, GPUs, quantum computers, neuromorphic systems). There is also a need to evaluate AI solutions for pattern recognition of events.
- Satellite observational data: The data need to be shared with a large community of scientists crossing multiple domains. The volume of the data presents interesting challenges in how it is managed and shared, but also in having access to the processing capabilities to analyze these large volumes of data.
- Decision support systems in health care: In 2018, the U.S spent ~\$3.65 trillion (18% of the GDP) on healthcare related expenses (up over 4% from 2017), and these are projected to grow at an average rate of 5.5% every year through 2027. Modern technological approaches to medical decision-making offer significant promise to an otherwise over-burdened and under-resourced healthcare system. Even small advances in health and medicine will impact quality of care for our entire population. Computational Health and Medicine is an emerging field that is critically dependent on efficient, secure access to diverse, geographically distributed health and health-related data: 1) How can we balance between secure access to private health information with the need for more diverse data for better informed health decision making? 2) What data models are well suited for individualized medical decision-making in a field weighted by one outcome versus other rare or more rare outcomes? 3) What algorithms are well suited to detecting rare outcomes in a field where the signal to noise ratio is very low?

Question 2 (maximum 600 words) – Data-Oriented CI Needed to Address the Research Question(s) and Challenge(s).

Considering the end-to-end scientific data-to-discovery (workflow) challenges, describe any limitations or absence of existing data-related CI capabilities and services, and/or specific technical and capacity advancements needed in data-related and other CI (e.g., advanced computing, data services, software infrastructure, applications, networking, cybersecurity) that must be addressed to accomplish the research question(s) and challenge(s) identified in Question 1. If possible, please also consider the required end-to-end structural, functional and performance characteristics for such CI services and capabilities. For instance, how can they respond to high levels of data heterogeneity, data integration and interoperability? To what degree can/should they be cross-disciplinary and domain-agnostic? What is required to promote ease of data discovery, publishing and access and delivery?

Providing resources, support, and other CI capabilities for data-intensive research is a complex challenge. Three types of researchers have been observed: (1) those with highly developed skills in porting their workflow and application(s) to new campus, regional, and national-scale resources, (2) those performing research on desktop systems discovering the need for scaling up to campus resources and beginning to realize what is possible on national-scale resources, and (3) those who know their limitations in desktop and campus computing and intend to scale up significantly and are beginning to understand what is available now and to anticipate innovations in the future. All three of these examples have cross-disciplinary or disciplinary-agnostic aspects and could include open research data and/or sensitive data (PHI, CUI, etc.) each with their own security requirements. Common challenges include: data movement from large instrumentation facilities to cyberinfrastructure sites; data movement from edge-computing and sensor devices to central computing/storage resources; data discovery across domains to identify available and relevant data; and cross-domain metadata translation. There is a great need to deploy advanced CI systems that lower the barriers to moving, analyzing, and sharing data. Important measures to achieve this goal include: seamless access through familiar web interfaces connected to the rich ecosystem of web services; on-demand interactive access to high performance systems; community data hosting services connected to high performance systems; and tools to easily manage diverse software environments. The broader community has, in part, responded to many of these challenges with a growing collection of domain and community specific data and discovery repositories. A “catalog of catalogs” could help the community easily find catalogs and data sets and possibly enable cross-catalog data discovery. This would require meta-data publishing and discovery services that will need to be integrated and available broadly to the entire community. Currently individual projects have to identify and adopt solutions on their own. With respect to storage services, although the cost of storage continues to plummet, high-performance file systems accessible from the supercomputers are often filled to capacity. This may be exacerbated by two factors: storing large files, especially those with low complexity, without applying compression and storing multiple copies of files. Deploying dedicated/specialized hardware for performing parallel data compression can address the former while software solutions for finding duplicate files can address the latter. Advancing the state of the art in decision support (particularly for health) relies on our ability to make sense of large stores of heterogeneous, incomplete, geographically distributed, protected data

sources. We must have the ability to compute and understand latent dimensions of risk and determine realizations of the quantities governed by them. We need the ability to get the data that traditional approaches to data science miss. To accomplish this, we need a workflow that simultaneously accommodates privacy, security, heterogeneity, performant computing, distributed communications, real-time interactive analysis, and scale. Underlying the challenges described above is security and privacy. In many instances the data is owned by one organization, the researcher is from a second organization, and the analysis resources the researcher wants to use is located at a third organization. Developing, articulating, and employing the processes for researchers to obtain the data, obtain access to and provision resources to perform the work, and pull this all together for use by a work group in many cases are complex and not well known to researchers. Performing this on campus resources, national resources and/or cloud resources should be considered in the solution(s). Many universities have not dealt successfully with these issues and having, documenting, and demonstrating one or more exemplars would serve the community well.

Question 3 (maximum 300 words) – Other considerations. Please discuss any other relevant aspects, such as organization, processes, learning and workforce development, access and sustainability, that need to be addressed; or any other issues more generally that NSF should consider.

People, organizations & communities are critical to the success of the National CI Ecosystem and should be central to any strategy put forward. Relying on loosely coordinated funding streams to address such a critical component of the ecosystem will limit impact to small, incremental gains. Maximizing the effectiveness of our existing workforce and the development of our future workforce depends on a coordinated funding strategy to make a significant investment in our only conduit to scientific and technological advancement. Workforce development programs in data science, embedded in undergraduate and graduate curricula in all disciplines of science and engineering, are required to efficiently harness the data revolution. A challenging consideration for many researchers is management of data following an award. Researchers often face inconsistent, impractical or limited guidance and requirements from publishers, funding agencies, domain communities, and their home institutions. Particularly in computationally and data-intensive work, the post-award preservation, maintenance, and support of data (particularly in the realms of volume and longevity) falls outside of the typical expectations or capabilities of many institutional or domain data repositories. A frequent issue raised by many researchers the XSEDE staff interact with, is that the needs for supporting data both during and after an award frequently face one fundamental issue: that lack of storage infrastructure on which this all must rest. Providing stable cyberinfrastructure services is essential. We must avoid disrupting the ability of the community to make productive use of current infrastructure, while addressing the need for innovative changes that will eventually accelerate scientific discovery. Innovative approaches should be developed alongside stable and tested solutions and deployed gradually. Lack of clarity regarding requirements or expectations for “reproducibility” in computational and data-intensive research. I.e., the possible assumption that reproducibility means bit-for-bit replication, as opposed to “scientific reproducibility,” or repeatability of a computational/data-intensive experiment.

Response to NSF 20-015, *Dear Colleague Letter: Request for Information on Data-Focused Cyberinfrastructure Needed to Support Future Data-Intensive Science and Engineering Research*

Reference ID: 11226901575_Towns

-- End Submission --