

Reference ID: 11226946393_Towns

Reference ID: 11226946393_Towns

Submission Date and Time: 12/16/2019 4:03:48 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 - National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign

Additional authors: Vikram Adve, Department of Computer Science, University of Illinois at Urbana-Champaign; Gabrielle Allen, Department of Astronomy, University of Illinois at Urbana-Champaign; Shawn T. Brown, Pittsburgh Supercomputing Center; Colleen Bushell, National Cent

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

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

Astronomy and astrophysics; cell and developmental biology; computer engineering; computer science; earth and environmental science; genomics; neuroinformatics.

Title of Response

Cyberinfrastructure for Instrument Science

Abstract

This letter focuses on two overriding issues for data-intensive science and engineering research: (1) being well prepared for the era in which advancing science requires diverse data from multiple instruments, such as LSST and WFIRST for astronomy, existing and emerging earth science data, anonymized health data supporting personalized medicine, and digital agriculture; and (2) enabling scientists to concentrate on the science, which requires a stable effort to provide state-of-the-art, interoperable data management in a manner that offers provenance sensitive to science needs. In the spirit of the successful NSF CI support for simulation science, a corresponding set of support must be developed for the distinct needs of instrumental science. These needs must consider proprietary data, high-velocity data, and the diversity of scientific instruments and measurements, including instrumentations producing a great volume of data. We foresee a common base of CI solutions, providing CI professional skills, software, and technical provisioning as a way to satisfy multi-domain instrumental science and to advance the state of the art for cyberinfrastructure that anticipates and accelerates use of CI advances by practicing scientists.

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.

Astronomy and Astrophysics: The primary research challenge for the astronomical community in the era of LSST and WFIRST is combining vast amounts of heterogeneous data from multiple missions that have previously been siloed, and using these combined data for discovery and inference. This is particularly crucial in the fields of multi-messenger astrophysics and discovery in the time-domain sky, which rely on signals over vastly different time and energy scales, from a disparate global network of observatories. Combining these datasets requires astrophysicists, statisticians, computer scientists and infrastructure from the commercial big data ecosystem to enable real-time and archival joint mission processing.

Biology and Health: The increasing resolution of scanning devices and vast ensembles of anonymized clinical data present new cyberinfrastructure challenges to the field of cancer research. In biology, projects such as the Earth Biogenome will capture tens of petabytes of genetic heritage of eukaryotic life before yet more diversity is lost. These data will enable scholarship in many areas, including bioengineering.

Digital Agriculture: Digital agriculture has created capabilities in data collection and information extraction that were unimaginable even a decade ago. Big data analytics, economic modeling, and forecasting will uncover relationships between local and global food security issues and resource management. Continued success in these areas critically depends on large interdisciplinary research teams, including computer scientists, engineers, and biologists with expertise in crops, animals, and microbes combining novel data collection methods with data sciences and machine learning to generate data on food production, markets, and food security outcomes. Datasets in agriculture are increasingly petabyte-scale, sometimes generated in a single day. Large-scale data storage and the capacity to compute across large datasets are important in many agriculture domains.

Earth Science: Key science investigations, such as those listed in the NASA 2014 Science Plan for Earth Sciences, require

an interdisciplinary approach to data arrays, open-source algorithms, and advanced computing capabilities. Indeed, these scientific investigations may well be the most important ones facing humankind. NASA alone holds ~100 PB of observational and modeling Earth science data, and there may be ~1 EB when combined with other Earth science data from other agencies nationally and internationally. In the next decade, these numbers will rise substantially. Advanced data capabilities related to machine learning, data engineering, and archival skills needed to manage exabyte scale data are needed, as well as technology upgrades, including connections to national and international datasets and computing facilities.

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?

Accelerating cutting-edge instrumental science increasingly requires efficient cyberinfrastructure (CI) oriented towards larger and more diverse instrumental data. NSF's strategy is to couple scientific discovery and CI innovation. Critically, there is an absence of CI capabilities and services optimized for extracting science from diverse instruments by modestly-funded investigators, including supplementing the skills typically found in small research groups with the cyberinfrastructure skills and capabilities optimized for extracting science from diverse instruments. A facility focused on coupling CI with instrumental science would provide an anchor to sustain ubiquitous access to the CI skills and software needed to advance the state-of-the-art in an analogous fashion to the CI NSF has developed to support simulation science. The facility would provide cross-domain observational data processing and data management methods, including: data engineering - realization of workflows, data packaging, and provenance, with special attention to calibrations and measurements; artificial intelligence - awareness of the evolving techniques of AI and machine learning that require vast amounts of structured data; data fusion - skills to map diverse data to a common schema, providing tractable access to data from diverse instrumental sources; virtual data - recomputation to trade-off persistent storage, especially for recomputations for comparison of successive processing runs; efficient data-intensive computation - hardware and software implementations that can efficiently move data in and out of databases and repositories with well-defined data-standards on the resulting computations so that derived data becomes as easy to query as primarily collected data from the instruments; curation and data lifecycle management - implementing decisions about lifetime of data, maintaining data in a viable format, and preserving any essential software functionality; curation and sustenance of open source algorithms; information security - effective practices for protecting the confidentiality, integrity,

and availability of data, which are especially needed in “cloud-like” environments where the investigators themselves carry responsibility; for example, protecting proprietary instrumental data; resource management - effective use of computational and data resources; for example, controls for runaway processing in pay-as-you-go environments; coupling between observation and simulation, for example, simulation of instruments to compare to actual measurements; and other topics related to volume, velocity, variety, and veracity of instrumental data sources. The supporting technology services would be distinct from MPI capabilities and services optimized for simulation science, which are typically renewed on five-year cadences. Contrasting this, instrumental science is often embarrassingly parallel, which allows for yearly incremental purchases, allowing for the facility to remain at the state of the art. The technology services needed for instrumental science include: throughput computing - coherent use of differing generations and speeds of processors; support for database engines, increasingly needed to support the necessary degree of data fusion; elastic user-controlled computing, supporting variable real-time data flows and computing demands; data lakes, facilitating ease of access to distributed and diverse instrumental datasets, including metadata indexing to support data discovery and enable enforcement of proprietary rights; excellent national and international connectivity to instruments and campus infrastructure; capability to preserve datasets beyond the period of active development; ability to provide allocations tuned to support multi-year grants. The collection of these resources must be easy to use and easily adoptable by modestly-sized scientific groups. This implies common management, best realized by a shared, federated facility, large enough to provide economies of scale, possessing the ability to evaluate tradeoffs between commercial clouds, provisioning at national centers, and provisioning at universities, consistent with the ability to move data as described in the NSF’s Blueprint for International R&E Network Connections.

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.

A facility would provide focus to bring to bear results of national grand challenge initiatives such as the National Quantum Initiative Act and the Executive Order on Maintaining American Leadership in Artificial Intelligence. These initiatives include both novel programming models and hardware architectures, and providing pathways for these novel techniques to become part of the working toolkits of practicing scientists. We expect a governance relationship between the facility, CI providers, and instrumental science that ensures the facility provides an agile, integrated, robust, trustworthy, and sustainable CI ecosystem coupled to the needs of instrumental science. The governance process will include the facility providing evidence that the resources are well-managed and facility users providing input on changing needs and gaps. Together, they will plan for staff and capabilities, and possibly facilitate the exchange of expertise, that best serve the community, given the resource constraints. Importantly, the CI facility will build a community of CI professionals and nurture a sustainable and appropriate technical workforce. Research processes producing data releases not only require scientific correctness but also must meet broad community expectations for digital data, enumerated by the FAIR principles. The underpinnings for findability, accessibility, interoperability, and reproducibility

reach into the research processes of the science groups. Providing data releases according to FAIR principles best relies on maximal use of common practices, well-supported by the CI facility we foresee. Formal datasets are typically the result of a release process that separates preparation and analysis from release. While the commercial cloud may be the place to deliver many formal datasets, preparation and analysis is the primary purpose of the facility we foresee. The facility would be organized optimally within the broader CI ecosystem with respect to commercial computing, computing at major facilities, and computing at universities, and would balance “pay-as-you-go” models with merit-awarded computing allocations.

-- End Submission --