

Reference ID: 11225491042_Purwanto

Reference ID: 11225491042_Purwanto

Submission Date and Time: 12/16/2019 7:34:47 AM

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: Wirawan Purwanto - Old Dominion University

Additional authors: Kent Carpenter; Old Dominion University; Dan Barshis; Old Dominion University; Eric Garcia; Old Dominion University; Ivan Lopez; Old Dominion University

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

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

High-performance computing; Bioinformatics; Bioinformatics; Bioinformatics; Bioinformatics

Title of Response

Observation of Challenges and Opportunities in Data-Intensive Sciences: Case of Bioinformatics

Abstract

Old Dominion University is a leading research university located in the southeastern Virginia city of Norfolk. ODU is home to a vibrant and diverse array of research endeavors in the areas such as cybersecurity, bioinformatics, chemistry, physics, computer vision, artificial intelligence, mechanical and

aerospace engineering, coastal resilience, business, and many others. Majority of these areas, if not all, have become data-intensive in recent years. ODU is strongly committed to investing in the cyberinfrastructure (CI) and support required by these research endeavors. This response represents a collective input from both domain researchers and CI provider at ODU. We present data-intensive challenges observed, as well as our suggestions for improvement that stemmed primarily from biological sciences. However, they are rather general and applicable to many other data-specific domains. The challenges can be summarized as follows: (1) challenge in storing and managing data given the increasing explosion of relevant research data; (2) data being siloed in disparate types of storage systems, which creates unnecessary friction in access, analysis, sharing, and collaboration involving these data; (3) the need for training to commensurately increase the skill level among CI users and practitioners to keep pace with the increased sophistication and complexity in data-intensive CI.

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.

Continual improvement of experimental techniques, collection techniques, and internet-of-things have led to rapid explosion of scientific research data, from which many invaluable insights can be mined. Drawing an example from biological sciences, current genome sequencing techniques allow us to see only a sample of an organism's genome, sometimes less than 0.1%, yet data generated per individual has already reached the sizes of many gigabytes and even terabytes. As sequencing techniques continue to improve, more genetic information will be captured, leading to even larger data sizes. These data become even more gargantuan as they are compared and combined to the RNA intermediates and the final protein products, which are goals in ecological and medical research targeting grand-challenge questions regarding life, diseases, etc. These lofty goals could be achieved only if scientists from an ever increasing number of domains collaborate, rather than working independently. At present, many researchers are using traditional high-performance computing (HPC) platforms to conduct research, which are still the most cost-effective due to the sheer amount of computing required. However, these platforms also lead to the fragmentation of data in silos (e.g. different HPC sites, different storage systems with varying complexity to store and access data). Limited storage often forces researchers to selectively keep data that were acquired with a lot of effort, or go through pains to transfer data from one location to another due to cost consideration. Data silos leads to unnecessary complexity and friction for data access by collaborators. Friction may include difficulty on bureaucratic level to gain access to data, technical (e.g. jumping over hoops to transfer data across two HPC sites), or financial (behind a paywall). Other challenges include: lack of widely accepted data management approaches/tools, common data quality and integrity control mechanism, proliferation of data formats, and challenges of (re)discovering data previously garnered. Many tools have been created by the scientific communities to address these challenges, yet they appear to be separate pieces which were development within only specific contexts rather than putting the holistic consideration of the data-

intensive ecosystem at the forefront. Prompted by the data-intensive needs in the commercial sectors, any companies have developed technologies in the cloud which appear to be promising to overcome many of these barriers; however, they are still new and not widely adopted in the scientific communities. Further, cloud financial model is still untenable, unpredictable, or unpalatable to many research entities.

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?

In an ideal world, there is a “one global data space” composed by datasets and knowledge garnered and shared by individual researchers. In addition, compute capabilities exist near the physical storage location of the data (in order to minimize transfer costs) which will enable scientists to perform the necessary analyses and computations with the relevant datasets. Scientists do not have to constantly worry about where the data resides (or whether they actually have a way to get access to the data), how to move data to the computation site, etc. Further, they need not be constantly bothered with low-level technical details on how to move data, etc. All these would happen in a seamless manner, for the most part behind the scenes, so that scientists would focus on the questions they have at hand. Future data-intensive CI platforms and capabilities should enable scientists to work with data on a high level: they would define the process and the workflow using high-level languages and scripts. Our current capabilities (e.g. the scientific software, file transfer tools, etc.) becomes a middleware by which these high-level commands would be translated into executable program invocations by the CI platforms as the “backend”. Future data-intensive CI should incorporate capabilities to address the following issues from its inception: * Friction-free pathway to data access, transfer, and sharing * Data management, quality, and integrity control * Data protection and access control * Standardized high-level API that are friendly to the end-user scientists * Supporting common data tools, usage patterns, approaches, algorithms that span across many domains The “friction-free” pathway would include, among others, high data transfer rates, commonly adopted APIs and connectors between storage types, geographic locations, characteristics. These challenges are not new and promising parts of the solution have been implemented by scientific communities. Examples include: GLOBUS and ScienceDMZ for secure, low-friction flow of data, iRODS for data management and secure, federated access to data across silos, science gateways, etc. Data-intensive capabilities, methods, and tools developed in the commercial sectors also need to be considered and adapted for scientific purposes. However, consensus and/or widespread adoption on the common approach/toolset to enable seamless collaborative research are still lacking. Further, these technologies are still present in “piecemeal” manner, waiting to be tightly

integrated into a single seamless platform. Many scientists and even CI providers are still unaware of these technologies, how to deploy and use them to provide seamless path for data access, sharing, and collaboration. For this to happen, translation of existing innovative tools and wider adoption of existing solutions need to be promoted. In addition, the development of higher-level tools/platforms leveraging the aforementioned tools and solutions need to be encouraged. Further, if cloud is to be part of this solution, then the question of its cost sustainability needs to be addressed.

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.

Beyond the above mentioned needs on CI resources, capabilities, and services, there is a great need for developing highly capable workforce to leverage the CI to carry out cutting-edge research. In many areas of endeavors, including biology and genomics, there is a great disparity in CI expertise, with the vast majority of researchers and students having a basic level of expertise, despite the ever wider adoption of the computational tools and capabilities in their research. For example, an average computational biologist tends to lack the awareness of the breadth of the CI solutions they can leverage, and they are not in a position to take advantage of the full power that CI resources, capabilities and services have to offer. Consequently, they are heavily dependent on publicly available programs/algorithms, and very few biologists have the training required to assess the quality of algorithms or capabilities on a particular research questions, improve existing capabilities, and/or develop new ones. Incentivizing additional training for and/or interdisciplinary partnerships between biologists, computer scientists, and/or computational scientists are crucial to make some progress on this front. Continued investment in workforce development will go a long way toward effective scientific work on the current and future CI. There is a clear trend towards more data and more computationally intensive analyses in biology, and the pace of this change is faster than the pace in which training is provided. Finding ways to scale up the training and increasing the general level of computing proficiency among domain scientists is a pressing need. Injecting CI training into college-level curricula will become essential to increase the baseline level of CI proficiency of future scientific workforce.

-- End Submission --