

Reference ID: 11223988641_Del Maestro

Reference ID: 11223988641_Del Maestro

Submission Date and Time: 12/15/2019 3:55:55 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: Adrian Del Maestro - University of Vermont

Additional authors: None

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

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

condensed matter physics; computational physics; quantum information physics; cold atom physics

Title of Response

Thoughts on Data-Focused Cyberinfrastructure Needed to Support Future Data-Intensive Science and Engineering

Abstract

Efficiently and accurately gaining insights from data intensive science and engineering will require investments in new cyberinfrastructure tools that aim to preserve and disseminate research products.

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.

As the director of an interdisciplinary research computing facility at a medium sized university, I observe daily challenges related to the ability to provide faculty with sufficient support to apply modern analysis methods (i.e. machine learning) to their domain specific data sets. They are often poorly trained themselves and are unable to find students that have the necessary skill sets required to work at the cutting edge. In my own research focused on algorithmic development for the quantum-many problem as applied to condensed matter and quantum information physics I see emerging challenges around the preservation and availability of trained deep neural network models. Increasing attempts to reproduce published results that employ machine learning in the physical sciences are time consuming and energetically wasteful as hyper-parameter optimization has to be continuously re-performed.

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?

The increasing availability of high precision experimental tools (e.g. neutron scattering) is producing a desire for improved simulation results, focused on larger and more physically realistic systems for comparison. This places increased burden on existing cyberinfrastructure as larger and more extensive simulations need to be performed producing large amounts of intermediary data that must be analyzed. Investment in cyberinfrastructure at research universities continues to fall below necessary levels while the administrative, application and reporting burdens of using these resources continues to increase. The lack of any broadly utilized repositories for data, models, and code that are centrally funded with prospects for long term support presents an increasing problem as researchers use the tools that are free and convenient (e.g. github). These were not constructed with science and engineering use cases in mind and may not even exist in the future. Increased investment should be made in developing tools that promote data, model, and code re-use. The discovery and utilization of these research products requires a broad and domain-agnostic approach to their acknowledgment and citation.

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.

Workforce development continues to be an issue as new tools and methodologies are introduced and their utilization rapidly becomes imperative to compete at the forefront of science and engineering. Dedicated support or funding models focused on research computing facilitation could have large impacts across diverse fields.

-- End Submission --