

Reference ID: 11221540213_Van Hooser

Reference ID: 11221540213_Van Hooser

Submission Date and Time: 12/13/2019 6:59:35 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: Stephen Van Hooser - Brandeis University

Additional authors: None

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

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

neuroscience

Title of Response

Data interfaces and databases for analysis in neurosciences

Abstract

More software is needed to organize, access, and share data, code, analyses, and analyses of analysis in the neurosciences

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.

At present in the interdisciplinary field of neuroscience, data is collected in individual labs in idiosyncratic ways. Each lab uses different blends of commercial or custom data acquisition and stimulus delivery machines that store data in a myriad of different formats. Further, the manner in which data is organized on disk differs greatly between labs and even within labs. Finally, new instruments and probes of different scales makes the development of static file formats difficult. This situation means that individual labs often need to write custom code to analyze their specific data, and that this code usually is specific to the data formats and manner of organization that is found in that lab (or even specific to an individual person). It can often be a month or more of work to analyze data from another lab. This lack of interchangeability greatly inhibits sharing of data and tools across labs, which silos data and ideas and reduces reliability and reduces development of gold standard tools. Interfaces are needed in order to programmatically address data acquired with different formats and organization.

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?

There are currently only a few attempts to develop standard data interfaces and standard data types. While one may say “how many are needed?” And dismiss the need to fund multiple efforts, the fact of the matter is that these standards have been slow to adopt because some of the approaches that have been tried are either difficult to use each time or have a high barrier to entry. So I would encourage the funding of many efforts in order that a few may gain traction. In addition, we need effective databases that can deal with storing raw data, analyses, analyses of analyses, and pipeline data, and to navigate the problems in research in which analyses may be controversial or buggy.

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.

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: 11221540213_Van Hooser

-- End Submission --