

Reference ID: 11227013023_Schneider

Reference ID: 11227013023_Schneider

Submission Date and Time: 12/16/2019 4:37:10 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: Nicole Schneider - Fulbright visiting researcher at Astronomical Observatory of Cagliari

Additional authors: None

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

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

computer science, machine learning

Title of Response

Need for Tools to Facilitate Collection and Analysis of Big Data

Abstract

Advancements in machine learning and data science are not always immediately available to domain experts in disciplines outside of computer science for a number of reasons. Inconsistencies in data formatting across disciplines makes it more difficult to spread novel techniques from one field to

another. Large datasets being collected require autonomous monitoring techniques that can detect when something has gone wrong with the survey, so that corrections can be made quickly. Lack of machine learning knowledge proves a barrier for scientists to try out complex machine learning techniques on their data. The solution to these challenges lies in developing infrastructure that will allow novel analytics to be performed on data from different disciplines by people with domain expertise rather than machine learning expertise.

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.

1. Different data formats make it difficult to do quick exploratory research across multiple domains, such as trying out a new machine learning algorithm on astronomy, geoscience, biology, and cyber security data to see for which applications it may be viable. Although there are some common formats, most disciplines store large datasets using a discipline-specific format. Some progress has been made on this issue, such as the National Institute of Standards and Technology's publication of the standard data format for election data. However, the challenge remains that when dealing with data from multiple disciplines there are not rigorous discipline-agnostic standard data formats. This makes it more difficult to spread novel ideas and techniques across domains, which impedes progress. 2. Large-scale surveys are becoming more common, and with this development come new challenges in terms of monitoring data collection. Large undertakings like the Large Synoptic Survey Telescope (LSST) need methods of monitoring the collection of data to ensure that problems that arise during collection (i.e. sensor failure, calibration problems, etc.) are caught and corrected in a timely manner to avoid the expense of recording meaningless or incorrect data. Human monitoring suffices for small scale operations but will prove infeasible for collecting data in the massive surveys that are becoming increasingly more common. 3. Machine learning is becoming a standard technique in many fields, particularly the sciences. However, the challenge remains that domain-specific knowledge is often required to tune models and assess whether the assumptions necessary to achieve meaningful results are met by the datasets being used. This means that extensive collaboration between data scientists and domain experts is often required, which leaves the open question: How can we facilitate the use of machine learning techniques by domain experts who do not have a machine learning background?

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?

1. To address the issue of domain-specific data formatting, common domain-agnostic standards can be developed for storing and making publicly available the massive datasets that are being collected through sensors and large-scale surveys. Alternatively, infrastructure can be developed to facilitate the conversion of large datasets from domain-specific to domain-agnostic formats, and to integrate datasets stored in different formats. Some work has been done towards standardizing, integrating, and making publicly available large astronomical datasets through the International Virtual Observatory Alliance which has a group dedicated to creating standards for virtual observatory remote data access. However, there is an absence of domain-agnostic versions of these standards. 2. To address the problem of efficiently monitoring data collection in large-scale surveys, infrastructure must be developed to automatically detect problems in the survey's hardware and software systems. Solutions may include domain agnostic real-time anomaly detection tools that learn patterns and alert when unusual events in the system occur. Infrastructure that takes advantage of parallel computing will be better suited to handle surveys that contain multiple sensors recording simultaneously. 3. One way to address the barriers that domain experts face when using machine learning on their data, is to develop infrastructure like applications to facilitate collaboration between data scientists and domain experts. For instance, platforms that allow scientists to test machine learning techniques on small samples of their data would be of use during the 'exploratory' phase of data analysis. Ideally, such platforms would allow public access to complex techniques, so that domain experts could apply them without needing to understand the underlying mechanisms of the techniques. Research groups should be able to add implementations of new techniques or algorithms to such platforms along with specifications of the assumptions that must be met in order to appropriately use them on a dataset. Further, tools could be included to assist users with discovering unusual patterns in their data, including mistakes like columns that contain only null values, anomalous datapoints, and invalid datatypes.

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.

NSF should consider funding curriculum development in high school courses in data science, and undergraduate courses in principles of Big Data collection and parallel computing. Further, basic data science training should be a fundamental requirement of undergraduate level applied sciences, engineering, and social sciences, because these disciplines frequently involve the collection and analysis of data.

-- End Submission --

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: «Respondent_ID»_«Primary_Last»
