

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: 11224180566\_Broude Geva

---

**Reference ID:** 11224180566\_Broude Geva

**Submission Date and Time:** 12/15/2019 7:20:42 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: Sharon Broude Geva - University of Michigan Office of Research

**Additional authors:** Brian Arbic; Daniel Atkins; Michael Cianfrocco; Oleg Gnedin; H.V. Jagadish; James Penner-Hahn; University of Michigan

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

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

Computational Oceanography and Climate and Space Sciences; Information; Electrical Engineering and Computer Science; Biological Chemistry; Astronomy; Chemistry and Biophysics;

**Title of Response**

University of Michigan Response to Request for Information on Data-Focused Cyberinfrastructure Needed to Support Future Data-Intensive Science and Engineering Research

**Abstract**

University of Michigan (U-M) research in computational and data-enabled science and engineering (CDSE) is highly interdisciplinary in nature and supported through active research programs in the university's 19 schools and colleges as well as large-scale centrally funded initiatives in computational and data science ([www.arc.umich.edu](http://www.arc.umich.edu)), mobility and transportation, exercise and sports science, precision health, etc. In the preparation of U-M's response to this RFI, CDSE faculty and researchers from across the institution were asked to comment on emerging data-intensive science and engineering research challenges particular to their field of inquiry. Due to the constraints on response-length, only representative input from faculty is included below to illustrate some of the more novel challenges reported. We expect the consumption of CI resources by U-M researchers to continue to expand significantly. Growth is expected across all disciplines, resource tiers, and usage modalities and would undoubtedly benefit from continued NSF investment in CI resources and services, programs to support hardware and software framework development, workforce development and curricular and non-curricular education programs specifically for CDSE, and additional guidelines and tools for research data management and use (from the end-to-end perspective, including policy and ethics considerations).

**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.

Structural biology of biological macromolecules using cryo-electron microscopy. Single-particle cryo-electron microscopy and cryo-electron tomography are used to study biological macromolecules. These techniques utilize transmission electron microscopes to acquire images of frozen biological samples by collecting high frame-rate movies for each exposure in the instrument. To overcome low signals, we collect thousands of movies for a single sample, resulting in typical datasets that are 5-10 TB. In 2020, U-M will have 4 instruments running 24/7, generating 20-40 TB per day for researchers across campus. We have engaged with 30 different labs across 4 units and 18 departments who would like to use this technology. (Michael Cianfrocco, Co-Director, Cryo-EM Center, Assistant Professor of Biological Chemistry, Research Assistant Professor, Life Sciences Institute; Assistant Professor of Biophysics) Physical/Inorganic Chemistry / soft condensed Physics. We're now using x-ray free-electron lasers (at Stanford, Osaka, and Hamburg) to measure excited state structures for various molecules. We collect data (currently at 100 Hz, but soon at 100's of kHz) that we need to move to Ann Arbor (with some level of pre-processing on-site but potentially involving megapixel images collected with MHz frequencies) and then process. (James Penner-Hahn, George A Lindsay Collegiate Professor of Chemistry and Biophysics) Physical oceanography, numerical ocean modeling, climate science. We perform simulations of global internal gravity waves that can only be performed at a few federal supercomputers. To take full advantage of these simulations, we need to store output at high frequency (E.g., every 10 minutes) over a few years, and at very small grid spacings in three spatial dimensions. Many in the oceanographic community would like access to the output, to plan field experiments and satellite missions, understand the interactions between high-frequency internal waves and lower-

frequency currents and eddies, and better understand how the internal waves break and drive mixing which affects marine biological productivity. Simulations of different models of this class exhibit different strengths and weaknesses, meaning that we would really like to have simulations of more than one model made easily available to the community. Another challenge is the effect that oceanic mesoscale eddies, the oceanic equivalent of weather systems, have on atmospheric weather and climate. We would like to have high-frequency output (every day, for instance) stored over long periods (up to a century), and at high spatial resolution. (Brian Arbic, Professor of Earth and Environmental Sciences and Professor of Climate and Space Sciences and Engineering)

**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?

On the needs for large data from instruments: For cryo-EM, instrumentation continues to be housed separately from data storage and compute. This contributes to a loss (or slowdown) of feedback from real-time data analysis considering the multiple steps required. Moreover, the disconnect of data from instrumentation leads to a loss of meta-data and the ability to analyze data in aggregate. This means that any future algorithmic improvements for cryo-EM image analysis will be slowed down as the data lives in multiple locations in a lab-specific manner, without standardized formats. Connecting these datasets to both short-term and long-term storage as well as to eventual compute locations is a major issue. Considering that the target audience is biochemists and biologists who aren't HPC experts, moving and analyzing this large amount of data is a major limitation. Web-based reproducible analysis frameworks that allow anyone with a web browser to access data and launch analysis routines would be a solution. Moreover, these workflows should be flexible, shareable, and customizable. There should be coordinated efforts across universities as this problem is faced by almost all US R1 universities (Cianfrocco) The main CI needs for our x-ray free-electron lasers projects are ways to transport large data files; potential need for large memory for data reduction, Compute-intensive processing to simulate data; data archival. (Penner-Hahn) On the needs for large model simulations: The CI needed to address oceanography questions is substantial. The ocean/atmosphere models described above can easily require tens of thousands of CPUs to simulate, over a period of months, and produce output ranging from hundreds of terabytes to tens of petabytes. The university-owned computer at Michigan serves the research community of an entire university and therefore cannot be set aside to serve just one research group. One solution, which my lab has gravitated towards, is to collaborate with scientists at federal labs, where substantial portions of the nation's largest computers can be set aside to solve the

sorts of problems mentioned above. Such collaborations require us to cede some control over the computations that are done because mission-driven agencies will satisfy their own priorities before they work on problems of interest to academic researchers. For the problems described above, the compute cycles required represent only part of the problem. One also has to contend with how to store hundreds of terabytes, or tens of petabytes, in perpetuity--and how to efficiently distribute these results to others in the community who may want access. Another problem is how to analyze such large datasets. For instance, a CS undergraduate intern in my group is still trying to finish a computation that he began during summer 2019. (Arbic) Current frontiers of cosmology and astrophysics are driven by flagship observational facilities and by diverse numerical simulations. The space observatories and giant ground-based observatories are so expensive that they require unified support from the entire community. They obtain data at the cutting edge of technology, which is by necessity incomplete. To fully utilize it requires supporting theoretical investigations. Current numerical simulations are designed and analyzed by relatively small research groups. Similarly to the observational facilities, future theory progress requires coordinated simulation efforts embraced, designed, and executed by all major groups, requiring novel solicitation formats that can provide umbrella support for multiple groups, on a scale much larger than previously attempted with TCAN. Flagship simulations would produce much larger amounts of data than is currently used, which should be analyzed by multiple groups. These data require universal repositories with modern database access (again similar to virtual observatories and archives) and dedicated human resources to support them. (Oleg Gnedin, Associate Professor of Astronomy)

**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.

Data Representation. In data science, we often re-purpose previously collected data. This greatly reduces costs, and make new insights possible. But it is critical to understand the coverage and representation of the dataset being used. E.g., major highways are in considerably better condition in Ohio than in Michigan. Therefore, vehicle-life data from one state may not be appropriate for the other even though they are neighbors with many economic similarities. (H.V. Jagadish, Bernard A Galler Collegiate Professor of Electrical Engineering and Computer Science, Director of the Michigan Institute for Data Science) Data Transparency. While it is well understood that one needs to record information about the manner of data collection, the variables recorded, etc., it is equally important to record the \*distribution\* of values for critical attributes relevant to the collected data. E.g., a dataset about industrial accidents may additionally need to indicate the company size distribution. We need enhanced metadata (or dataset label) association techniques. Furthermore, these annotations should be preserved and propagated through transformations, (Jagadish) The role of AI /ML and automation, broadly speaking, in data management and research workflows. Workflow management systems and tools were developed to address “productivity” and error-avoidance needs, to support reproducibility and replicability. Efforts such as Project Jupyter and the Open Science Framework have attracted a growing user base in experimental sciences characterized by smaller data sets, but there is a growing interest in applying AI/ML approaches to a range of research tasks. Recent advances in these areas offer

*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: 11224180566\_Broude Geva

---

the opportunity to reimagine research workflows in ways that can vastly increase the volume and efficiency of scientific research and improve research outcomes. Among other aspects, policy and ethics need to be part of the discussion. (Daniel Atkins, Professor Emeritus of Information, Professor Emeritus of Electrical Engineering and Computer Science)

-- End Submission --