

Reference ID: 11226910335\_Morgan

---

**Reference ID:** 11226910335\_Morgan

**Submission Date and Time:** 12/16/2019 3:49:43 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: Dane Morgan - University of Wisconsin, Madison

**Additional authors:** None

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

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

Materials science; materials informatics; computational materials science

#### **Title of Response**

Materials Data and Machine Learning Cyberinfrastructure

#### **Abstract**

Infrastructure to support machine learning model generation and dissemination would enhance the impact of this rapidly developing area.

**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.

A key challenge facing S&E researchers is obtaining materials data, and we increasingly need to have materials data that is accessible by computer codes in an automated fashion. This includes such data as basic properties in databases, which is already being actively shared in many ways, and also data on synthesis, processing, and performance. This data is important for traditional approaches to MS&E, and will become increasingly important as more activities are automated in code and through artificial intelligence.

**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?

To support the data challenges above machine learning (ML) offers exciting new opportunities. The community is embracing ML in the context of generating many new models, but there still seems to be very limited reuse of such models to actually solve materials design problems. There is a need for more CI and associated cultural changes to support use of ML models as they are developed. In particular, CI that supports easy development and dissemination of quality vetted ML models would be extremely valuable. This should include the ability to interact with trained models easily through APIs, integrate multiple models, and improved/evolve models easily with proper citations and provenance tracking.

**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.

There is both a need and an opportunity to engage undergraduate students from MSE in data science activities. This engagement will increase their specific skills and their general data science literacy. This engagement can happen in classes but hands on activities are essential, e.g., through class or outside-of-class research projects. Furthermore, the excitement around AI/ML can attract students who might otherwise not have explored research. We have been exploring undergraduate research at the interface

*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: 11226910335\_Morgan

---

of data and materials science at UW through the Informatics Skunkworks (<https://skunkworks.engr.wisc.edu/>) and found excellent engagement from students.

-- End Submission --