

Reference ID: 11224169538\_Voyles

---

**Reference ID:** 11224169538\_Voyles

**Submission Date and Time:** 12/15/2019 7:09:24 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: Paul Voyles - Wisconsin MRSEC and Department of Materials Science and Engineering, University of Wisconsin-Madison

**Additional authors:** Victor Zavala, Wisconsin MRSEC and Department of Chemical and Biological Engineering, University of Wisconsin-Madison; Dane Morgan, Wisconsin MRSEC and Department of Materials Science and Engineering, University of Wisconsin-Madison

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

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

Materials science, chemical engineering, structure of materials, materials informatics, computational materials science, sensing

### **Title of Response**

Challenges and Opportunities in Data-Intensive Research in Materials Science and Engineering

### **Abstract**

Researchers in materials science and engineering increasingly view data, databases, and machine learning as a co-equal toolset with experiments and physics-based simulations. We identify three key challenges to future data-intensive research in materials: (1) Aggregating data on materials across various data sources, then integrating the data with expert knowledge using machine learning, all to design new materials. (2) Extracting materials data from heterogeneous, natural language sources such as the primary research literature. (3) Deriving materials information from large-scale data produced by current and next-generation materials characterization experiments, including co-design of instrumentation and algorithms. To meet these challenges, we suggest two forms of data-intensive cyberinfrastructure. (1) A software and hardware ecosystem for development, dissemination, retraining, and reuse of machine learning models in materials and beyond. (2) An integrated data infrastructure for materials characterization spanning smaller-scale computing dedicated to specific instruments up to facility-scale computing like XSEDE, enabling real-time, near real-time, and offline analysis. Overall, cyberinfrastructure should be paired with and informed by a cross-cutting effort to develop, disseminate, and instantiate in software best practices for data and machine learning in materials and cross-disciplinary training of materials and machine learning domain specialists.

**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. Obtaining values of materials properties for candidate materials to aid materials design. Example properties include band gap, elastic constants, and thermodynamic stability. Example design problems include improved photovoltaic absorbers, recyclable polymers, and materials for extreme environments. Currently, property values are sometimes available in open databases with good APIs, but more often they can only be found in limited access databases, review papers, or isolated studies in the literature. Property values also may be available indirectly from related data correlated with the property, from physical models such as DFT, CALPHAD, and others, or increasingly from data-driven machine learning models. Researchers need the ability to easily aggregate and integrate data from all these sources and then use that data with their expert knowledge in machine learning algorithms. Existing algorithms are not suitable for this task, which often results in data from disparate sources being incorporated in an ad hoc manner. Algorithms that support incorporating expert knowledge are critical to reduce the amount of data needed and to enable suitable generalizations. Highly interoperable resources with streamlined access that aggregate data and predictive models would be valuable for the community. 2. Extracting materials information from heterogeneous sources of data. The field needs systems for automatic extraction of information from the scientific literature and other texts, including all elements of the materials tetrahedron (synthesis/processing information, properties, structure, and performance). This ability would dramatically accelerate the rate at which researchers can aggregate and apply information, for example on how to synthesize new compounds or identify materials with target properties. This challenge is a special case of the general problem of natural language processing in science. 3.

Understanding materials structure, properties, and processes from large-scale image data from materials characterization. Current generation synchrotron beamlines, electron microscopes, and scanned probe microscopes offer unprecedented views of materials structure, functional properties, and processes such as phase transformations and ferroic switching. They also routinely create terabytes of data, and next generation even more capable experiments will produce even larger datasets. Automated analysis is required to derive materials information from these data with limited human intervention, including robust anomaly detection to allow for serendipitous discoveries. Development of automated analysis also will create new opportunities for co-design of next-generation hardware intimately merged with and enabled by software in a computational imaging paradigm. This challenge is a special case of the general problems of computer vision and computational imaging.

**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?

Cyberinfrastructure in the form of an ecosystem to support easy development, sharing, use, assessing, retraining, and integrating machine learning models is needed. We will increasingly be codifying our data and knowledge in the form of machine learning and physical models, and future material design activities will need to be able to interact with such models instantly through the cloud with powerful APIs. Such robust ecosystem of machine learning also must support the maintenance of a models as evolving, fully accessible cloud resources, not just static results in a paper or even a Github repository. The models in the ecosystem need to become more like Wikipedia, constantly evolving to reflect our consensus understanding, and less like a traditional encyclopedia, which quickly becomes out of date. Achieving this capability will require a strong coupling of software development platforms (such as Github), cloud access, machine learning software (such as Pytorch), coupling to existing and user data, and flexibility for use and modification. Cyberinfrastructure of materials characterization data can be defined in terms of the turnaround time for analysis. Real-time analysis is required for immediate feedback to ongoing experiments, including self-guiding, autonomous experiments capable, for example, of selecting the right data to acquire next based on analysis of the data acquired so far. Near real-time analysis taking ~10 minutes or less is required to enable fine tuning of experiments during an experimental session, either by human operators or algorithms. Off-line analysis requiring days or weeks of wall-clock time for computing is acceptable for post-experiment analysis. Deploying all of these capabilities requires an integrated, data-centric computing ecosystem which integrates real-time or near real-time data reduction and analysis with off-line, computationally expensive analysis. Larger-scale

computing resources are often available at the institutional level or through resources like NSF XSEDE, but networking bottlenecks for transfer of large data sets create a need for additional dedicated computing near large data producing instruments. However, the local computing needs to be seamlessly connected to larger-scale resources.

**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 a need for cross-training of materials and ML domain specialists, which is shared by many domains of NSF-sponsored research. In materials science specifically, there is a need for scientists and engineers with skills for efficient problem solving and research using modern cyberinfrastructure, such as image analysis, ML, Github tools, Jupyter notebooks, automated computing workflows, and physics-based simulations. Materials science and materials scientists are dispersed across a wide variety of disciplinary homes and related professional societies, including TMS, MRS, ASM, APS, ACS, ACerS, and others. There is a need for coordination across these boundaries and amongst these organization on standards and best practices for disclosure, use, re-use, and appropriate credit for data and machine learning models. These best practices should be support by software tools, potentially associated with the modeling and analysis cyberinfrastructure mentioned above.

-- End Submission --