

Reference ID: 11227063185\_Shankar

---

**Reference ID:** 11227063185\_Shankar

**Submission Date and Time:** 12/16/2019 4:51:19 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: Sadasivan Shankar - Harvard University

**Additional authors:** G. Crabtree (Argonne); R. Eggert (Critical Materials Institute); T. Lograsso (Critical Materials Institute), J. Neaton (Lawrence Berkeley Lab, UC-Berkeley), D. Prendergast (Lawrence Berkeley Lab), B. Sumpter (Oak Ridge National Laboratory), S. Whitelam (L

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

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

Materials Science and Engineering; Chemistry; Computing and Information Processing; Machine Learning for Scientific and Engineering Applications

### **Title of Response**

An Open Cyberinfrastructure for Data-based Research in Materials and Chemical Sciences

### **Abstract**

Our intent in this document is to highlight research and application challenges in building and sustaining a cyberinfrastructure initiative in the US. The two sides together could enable a multiplicative effort for research in chemistry and materials science and engineering. This infrastructure could help modelers and experimentalists including ability to tools to focus experimental efforts to provide limited data where it is most needed within a much larger space. As the authors have been exposed to use of data for both scientific and technology breakthroughs, in academia, national laboratories and industry, our perspectives have been shaped by working together with large teams to deliver solutions. For the US to sustain the leadership in this Machine-led century, it should explore expansively the vision for how to bring communities together. It is important that such an initiative should not encompass nucleation, but also should be made sustainable long term. There are lessons to be learned from the Silicon Valley and elsewhere of how this can be addressed. In addition, the communities of scientists and engineers from academia, national laboratories, and industries should feel inclusive and empowered by this effort. This implies including in addition to the usual research aspects, sustainable aspects such as security, intellectual property prediction, credit for sharing data, long term funding viability.

**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.

Data has been considered as the fourth pillar on which scientific and engineering research have been built upon. This has been made possible due to the relentless evolution of Moore’s law over the last fifty years and all the inventions built on top of it. Mining and analyzing data with advanced statistical and computing techniques aid in uncovering basic principles that are hard to isolate otherwise. The utility and applicability to materials and chemistry are immense. It is expected that engineering and design of new materials and chemical molecules will be enabled. These are due to the complexity of designing materials for real applications. The key research questions related to data-based analysis are the following: 1) For already existing data, how can you determine the statistics? 2) How to transfer data from different sources and curate it so that users can access the data? This is a critical problem and is caused due to the following reasons: a) Experimental setups are different from simulations-generated data; b) All data are in relation to a specific application, equipment, and analytical tools. 3) Since data-based analysis reveals correlation, how would one connect with further analysis to determine causation? How do we build physical hypothesis into our data analysis tools? Related to this is the requirement that data-based analysis is not a substitute for experimental or computational methods and they all need to be complementary. The applicability of data-based analysis is gated by the following challenges: 4) Availability of data with known statistical variations; 5) Annotations or meta-data associated with raw data; 6) Infrastructure to access, transfer, and process large amounts of data with different statistics and sources; 7) Security for the data, especially different national laboratories (e.g Department of Energy) and industrial partners, which needs to be traded-off with open science guidelines; 8) Scaling of “big data” analytic methods to “small data” problems in materials and

chemical sciences and engineering, with some combination of physical modeling; 9) Visualization of data and the associated meta-data.

**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?

The advent of data as a new paradigm has been enabled by the Moore's law and its consequences, including the availability of large quantities of data associated with the internet and the new algorithms. The research challenges are elaborated below: 1) Statistics of Existing Data: This is important as there are large volumes of data especially from different experimental set-ups, especially in national laboratories in their user facilities. For the older data to have practical value, the statistics need to be well bounded. 2) Transferability of data between different sources: Transferring data from different sources is critical for users to access similar data or be able to use workflows as appropriate. This is a one of the biggest problems given the differences between the sources and the organizational methods of storing data. It is one key for developing the AI and machine learning capabilities for the upcoming decades. 3) Causation and Correlation: Since data-based analysis reveals correlation, it is important to differentiate it from causation. Previously, we had provided several guiding principles to use data-based analytical methods for applications to sciences and engineering. Data-based analysis is not a substitute for experimental or computational methods, but are complementary. More caution and care are needed to use these methods, especially given the hype of data-based methods as panacea in all fields of research. Tracking data provenance can potentially help identify what/were the experimental differences led to incompatible data. Whether it was because of sample misalignment, or differences in sample preparation that were overseen, etc., to ensure high quality research. 4) Statistical Variations of new data: The data resulting from large simulations are typically expected to be precise. However, the simulation methods themselves use various numerical approximations both in the inputs and also in the models. This needs to be studied as part of the research problems and also should be available with the data, or as meta-data associated with the given data. 5) Meta-data associated with raw data: Data resulting from both experimental measurements or computer simulations are driven by specific applications. Associated with each measurement are the specific instrument-related calibration and associated parameters. Associated with each simulation are modeling parameters and bounds on validity. Since both the instruments and simulation models change with time, it is important that meta-data be made available with the raw data. This should take into account intellectual property constraints of tool-associated meta-data. We

would also suggest learning from other existing platforms that leverage editable, but controlled open frameworks that balance multiple constraints. 6) Infrastructure for large amounts of data: For a community-level leverage of data-based analysis, a cyber infrastructure should have the following attributes: 1) Secure access of the data by multiple users; 2) Secure transfer of the data between collaborators; 3) Ability to process data by using open or custom machine learning tools. Collaborating with national laboratories may help address this aspect. 7) Security: Security of data is critical especially different national laboratories and possible industrial partners. The security needs to be traded-off with open science requirements. 8) Scaling of “big data” analytics to “small data” in materials/chemical sciences and engineering: Many of the problems in materials and chemical sciences/engineering do not generate large volumes of similar kinds of data. As a result, data-based methods need to be appropriately reformulated and scaled for smaller data from multiple sources. 9) Visualization of data and the associated meta-data: Data visualization need to be supplemented with the associated meta-data for proper analysis. In addition, most of the chemical and materials problems relatively are on larger dimensions. Intellectual properties associated with commercial tools need to be resolved.

**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.

**Organizational Outreach:** Data availability has exploded driven by availability of hardware, instrumentation, and availability of clouds to store both public and private data. NSF has funded many centers and efforts which are potential sources of data for materials and chemical sciences. There are substantial efforts in Department of Energy labs and research hubs which collect vast amounts of data. Examples include Critical Materials Institute, Joint Center for Energy Storage Research, and large national labs such as Argonne, Lawrence Berkeley, Oak Ridge, etc. A common method of collecting data to the different entities may be difficult from a practical perspective. However, developing an infrastructure should involve discussions with the different labs to ensure that the platform is usable across organizations, without necessarily making all the data open. Our preliminary communications have revealed that the researchers in different DOE centers are open to discussions. In addition, inputs from industrial partners will be useful in this regard. **Processes:** It is important that any infrastructure development should be both forward looking using lessons from the past. In addition, the best practiced methods from different organizations need to be shared without the necessity of data sharing. **Workforce Development:** It is important that data science should be integrated with the regular curriculum or as part of laboratory training. If the infrastructure is available and does not add large overheads to research, scientists and engineers will be able to see the value of being part of a larger effort on data-based analysis. **Sustainability:** Although there have been very successful efforts in the US for scientific computing, which have led to several new tools, we have always lagged behind in a long-term support model. This is another area that collaboration with the Department of Energy efforts would be valuable as universities have difficulties in maintenance of codes and infrastructures.

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: 11227063185\_Shankar

---

-- End Submission --