

Reference ID: 11226931034\_Atherton

---

**Reference ID:** 11226931034\_Atherton

**Submission Date and Time:** 12/16/2019 3:57:45 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: Timothy Atherton - Tufts University

**Additional authors:** None

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

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

Condensed Matter and Materials Theory; Soft Matter

#### **Title of Response**

Cyberinfrastructure for Soft Matter

#### **Abstract**

Soft matter research could be highly catalyzed by CI investment because the field naturally involves highly disparate datasets, challenging machine vision problems and highly complex simulation tasks. Development of a centralized data repository as integrated simulation and analysis tools should be considered high priorities for NSF investment in Cyberinfrastructure. These must be accompanied by

appropriate training opportunities as well as community driven design to ensure maximum benefit to the user community.

**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.

The Soft Matter field aims to discover universal principles that link seemingly disparate media and use these to create new materials with desired properties. Several exciting and interlinked themes that have recently emerged from these efforts include: 1) Geometry and topology as design principles, including complex and curved geometries and new types of order, defects, etc.; 2) Composite materials and metamaterials with extreme mechanical properties; 3) “Active” materials that continuously dissipate energy including swimmers, swarms and biological matter; 4) Shapeshifting materials including soft robotics and elastomers. Theoretical and experimental developments in each of these areas are highly interlinked but the field has a number of unmet cyberinfrastructure needs that, if supported by the NSF could greatly enhance progress in this field.

**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) Much work in Soft Matter fundamentally relies on quantitative analysis of image data obtained by many different kinds of microscopy, and involves extraction of correlation functions, order parameters, etc. This work is often done manually at present, and cyberinfrastructure tools readily usable by non-experts would immeasurably benefit these efforts. Integration with experimental apparatus to enable real time and high volume analysis would greatly benefit experimentalists and enable much more robust tests of theory. Particularly exciting is that many of these tasks fall into the domain of categorization and labeling problems that are now efficiently performed by Machine Learning. A well-documented Soft Matter oriented ML package with tutorials might be particularly valuable, but it is essential that it be high performance. 2) Computational approaches to Soft Matter often rely on simulation. While there are some very high quality open source packages

(particularly for techniques like Molecular Dynamics), the needs of soft matter researchers often push beyond the capabilities of commercially available tools (e.g. COMSOL for finite elements). As a result, the field has many homegrown codes that are at best only weakly available to other researchers, and little thought has gone into their wider use or interoperability. In addition to funding development of further, well-documented and general purpose codes, it would be extremely helpful for NSF to sponsor workshops that drive better integration. Collecting soft matter researchers together with CS/CI experts and designing data formats, integration opportunities etc would likely pay large dividends as this sort of thing has never happened before. 3) Soft Matter lacks a central repository for published data. As a result, many valuable datasets produced in the community are inaccessible and reanalysis of data as techniques are developed is challenging or impossible. In part, this limitation arises because of the extreme heterogeneity of the data. This is in contrast to highly successful efforts in close fields e.g. the Protein Data Bank. Development and maintenance of such a repository could be highly transformative, but requires a community-led design process (interrelated with (2) above).

**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.

A particular challenge is that Soft Matter users of CI are often not experts in computing, and hence adoption is limited by training opportunities. These could take the form of online manuals, tutorials, but workshops where participants work with their actual data should be an essential component of CI development. As I discuss above, interoperability requirements and data heterogeneity inherent to the field means that a community driven design process is essential so that CI are widely beneficial.

-- End Submission --