

Reference ID: 11226974616\_von Oehsen

---

**Reference ID:** 11226974616\_von Oehsen

**Submission Date and Time:** 12/16/2019 4:14:54 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: Barr von Oehsen - Rutgers University

**Additional authors:** John Barden, Yale University; Rene Baston, Columbia University/Northeast Big Data Innovation Hub; John Goodhue, MGHPCC; Vasant Honavar, Pennsylvania State University; Wendy Huntoon, KINBER; David Marble, OSHEAN; John Moore, Consultant; Sharon Pitt, Univer

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

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

University CIOs; Research and Education Network Presidents; University Research Computing Directors; Research Faculty; Center Executive Director; Associate Vice President

### **Title of Response**

The Eastern Regional Network (ERN) Multi-Campus Federated Data-Focused Cyberinfrastructure in support of Data-Intensive Science and Engineering Research Collaborations

### **Abstract**

The Eastern Regional Network (ERN) was formed to more effectively address challenges related to simplifying multi-campus collaborations and partnerships in the Northeast that advance the frontiers of research, pedagogy, and innovation. This vision and mission reflects the reality that multi-institutional collaborations are on the rise, but the data sets that support them are getting too large to transfer easily, the data-intensive computing resources that they require have increased beyond the capacity of most campuses, and the expertise needed to support data-intensive science is scarce. The now ubiquitous role of research computing and data in scientific discovery and scholarship across all disciplines presents new challenges not only to providers of these services but also to researchers eager to keep pace with the ever-changing technology landscape. Even more challenging are multi-institutional/multi-investigator research projects that require shared access to low-latency or high-bandwidth networks, data repositories, advanced computing resources, and specialized research instruments. To address these challenges, the ERN requires funding for Multi-Campus Federated Data-Focused Cyberinfrastructure in order to successfully support Data-Intensive Science and Engineering Research Collaborations. Having this as an NSF priority will allow us to work towards building a cohesive data-centric ecosystem that spans the Northeast and beyond.

**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.

Cryo-electron Microscopy (cryo-EM) is a new method for determining the structure of biomolecules that has revolutionized structural biology. More recently, Electron Cryotomography (Cryo-ET) is using similar technology to expand the range of available imaging results. While Cryo-EM and Cryo-ET owe much of their power to advances in microscope optics and detectors, they are equally dependent on image processing pipelines that are both compute and data intensive. Exploratory conversations with Cryo-EM/Cryo-ET leadership at Rutgers, Yale, the University of Massachusetts, Penn State, and others suggest that a stronger partnership between providers of Cryo-EM and cyberinfrastructure and services can both improve the efficiency with which well-resourced labs can obtain scientific results and make this technique more readily available to underserved institutions that have fewer resources and less access to technical expertise. Materials Discovery is an area where recent advances in machine learning, together with advances in computing and high throughput measurement techniques offer the potential for new data-driven approaches to predicting properties of materials based on their descriptors. When suitable data exists or can be generated, these methods can be useful to determine material properties that are too expensive or time-consuming to measure or compute using traditional methods. The descriptors may be of many types and scales, depending on the application domain and needs. Predictions may be interpolative or extrapolative, allowing the design of entirely new materials. Conversations with Penn State, Rutgers, SUNY Buffalo, MIT, and others suggest that Materials Discovery offers an attractive testbed for advanced cyberinfrastructure of the sort the ERN can offer to facilitate data and computation intensive research. We are excited about the Cryo-EM and Material Discovery

initiatives that are starting now, which will build on our initial cyberinfrastructure collaborations and should pave the way for an abundance of science and education outcomes in a range of disciplines. Over time, these efforts will have transformative impact on the research and education communities in both small and large institutions across the region. Having access to funding will help speed the process of developing a distributed federated data-intensive cyberinfrastructure ecosystem that will directly benefit both the Cryo-EM and Materials Discovery communities. It will also lay the groundwork for future collaborations that benefit other data-driven science communities.

**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?

We see three significant sets of challenges. The first two are common across the national research community, and the third is unique to the Northeast. The now ubiquitous role of technology and data in scientific discovery across all disciplines, including Cryo-EM and Materials Discovery, has created new challenges as providers of these services try to keep pace with the ever-changing technology landscape and the demand from the research community for new and emerging technologies. To make this even more challenging, many research projects are now span multiple campuses and organizations, requiring access to low-latency and/or high-bandwidth networks, distributed data repositories, and shared computing resources. Added to this is the demand for infrastructure designed around reproducibility and containerization of applications and workflows. Traditional solutions such as university condominium advanced computing models, enterprise network infrastructure, nationally funded initiatives (XSEDE, OSG, GENI, CloudLab, Chameleon), and to some extent commercial clouds are not able to offer the flexibility, ease of use, and access to the technologies (especially new and emerging) needed to keep researchers competitive their fields of expertise or to support collaborative projects distributed locally, regionally, and/or nationally. The advent of Big Data and collaborative research requires an approach to research computing and data science that facilitates the analysis and sharing of ever-growing data sets by virtually all fields of scholarship. Access to a regionally distributed federated resources, complemented by diverse support personal to aid faculty, students, and staff, will go a long way toward meeting the evolving needs of current, new, and emerging research and education communities. Funding that will allow us to federate and leverage existing and future services, data, and people from multiple campuses, organizations, and regional network providers is essential to success. Another important challenge is the need to ensure compliance with standards for protecting sensitive information (e.g. NIST 800-171, dbGaP, HIPAA, and FISMA). While well-understood techniques

for protecting sensitive data exist, establishing efficient institution-wide controls that satisfy the requirements of numerous overlapping standards without unreasonable burdens on individual researchers is a non-trivial challenge. Addressing these challenges is critical to the success of NSF programs such as CC\*, Harnessing the Data Revolution, or CSSI is necessary. As the ERN looks for ways to simplify research collaborations and sharing of resources in the Northeast, we are faced with historical and emerging challenges that include:

- End-to-end data transfer/access requires working with multiple regional research and education network (REN) providers (at least 9 within the region);
- The density and diversity of colleges and universities in the region -- more than 1,930 public and private institutions in many shapes and sizes, in both EPSCoR and non-EPSCoR jurisdictions
- The states in the region are smaller than states in the West—meaning that most multi-institutional collaborations cross state borders
- The Northeast contains eight different state university systems in a geographic area whose size is comparable to that of California
- The Northeast has many colleges and universities that have existed for more than a century. While they can be important sources of strength and stability, they can also be slow to change.

**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.

The Eastern Regional Network (ERN) was formed to address challenges related to simplifying multi-campus collaborations and partnerships in the Northeast that advance the frontiers of research, pedagogy, and innovation. The ERN is first and foremost a network of people interested in pursuing this vision, and who manage and use campus and regional research computing, data, storage and network resources that can make it happen. The ERN's mission is to provide layered and transparent access, to federated data and computing facilities for research projects located at and between partner sites. The resulting research and education platform will support a diverse set of science drivers and emergent educational opportunities and offer the research community access to a broad range of collaborative multi-institutional resources that are not available on any one campus alone. The vision and mission of the ERN reflects the reality that multi-institutional collaborations are on the rise, the data sets that support them are getting too large to transfer easily, the computing resources that they require have increased beyond the capacity of most campuses, and the expertise needed to support compute intensive research is scarce. To address these challenges, the ERN focuses on the whole stack - organization, processes, learning and workforce development, access and sustainability. We are excited about many of the data intensive research initiatives that are starting now, which will help the ERN community build federated infrastructure that facilitate collaborations that will one day pave the way for an abundance of science and education outcomes in a range of disciplines. Over time, these efforts will have transformative impact on the research and education communities in both small and large institutions across the region.

-- End Submission --

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: «Respondent\_ID»\_«Primary\_Last»

---