

Reference ID: 11221000594\_Peckham

---

**Reference ID:** 11221000594\_Peckham

**Submission Date and Time:** 12/13/2019 2:33:09 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: Joan Peckham - University of Rhode Island

**Additional authors:** Christopher Hemme, University of Rhode Island

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

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

Computer Science, Data Science, Bioinformatics

#### **Title of Response**

Community support for Data and Data Driven Computing

#### **Abstract**

We have read the NSF Blueprint for Data-Focused National Cyberinfrastructure Coordination Services, and resonate with attention to agility, interdisciplinary engagement, and the development of domain agnostic and general purpose infrastructure. Here we outline our needs for building inter- and intra-institutional networks for support of rapidly evolving scholarly discoveries across domains and

institutions. Our area of highest priority is the development of support for connections, training, and collaborative communities of scholars in need of computational and data infrastructure. We need help to further develop an ecosystem that renders emerging technologies understandable and accessible to domain experts, and helps technologists to understand the urgent needs of domain experts.

**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.

Research Challenges: Similar to other institutions, Data Science and Big Data is interdisciplinary here in that most data science scholarship is driven by problems in one or more domains, but then is applied to several others, becoming domain agnostic. Examples of data science research and scholarship at our institution include:

- Algorithms for efficiently searching very large data sets by taking advantage of identifiable geometric and topological characteristics (initially astronomical and biological, but extensible to multiple other disciplines including oceanography and ecological science)
- 

Machine and deep learning techniques for structured data, initially developed for computer vision, computer program analysis and computational biology, but now applied to cybersecurity data, for example)

- Modern statistical methods that require computational approaches including Bayesian, network, survival, and longitudinal statistics; missing data techniques; causal inference; multivariate analysis; dimension reduction; variable selection; empirical likelihood. These computational strategies are applied to environmental sciences, medicine, spatial epidemiology, electronic health records, and finance, for example

Computational and statistical scholars are working with partners across the colleges in multiple domains, developing techniques to solve data driven problems in specific domains, and then developing agnostic techniques to apply to multiple other domains where appropriate. For example at our institution, data analysis to determine the impact of climate change on coastal regions (NSF EPSCoR award), biomedical research (NIH INBRE award), geophysical seismological oceanography (several past and recent NSF research awards), genomic and functional diversity of oceanographic bacteria, collaborative research using artificial intelligence techniques to understand memory in neuronal networks (recent OAC/CISE award), nutritional epidemiology, influence of biochemical capabilities of microbial communities on biogeochemical cycles and food webs in aquatic environments (investigator with joint position in Oceanography and Cell & Molecular Biology and a co-PI for our NSF EPSCoR award).

**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?

Below are a few places where national, regional and local resources do not yet fulfill need.

- New and emerging technologies and strategies, including data pipelines and workflows, GPU and HPU computing, artificial intelligence, cloud computing, security/privacy and ethics. In the future we will need assistance for translation of the newest technologies and approaches to multiple domains, for example quantum computing. We note that while scientists were the first to seek help with data, our partners from the social sciences and humanities are also now signaling the need for technical data infrastructure.
- Shortage of viable data storage and archival concerns us. Robust or standard means to publish data in support of the scientific method, which calls for availability of data for reproducibility. We do not have robust and trustworthy systems capable of archiving important data, metadata, and provenance in standard ways that assure security but permit sharing and support reproducible results. This includes archival solutions that are secure and resistant to climate or other disasters.
- Automated system tools. For example, tools that could assess a program to determine how it is spending its time between I/O and computation in the HPC environment and then automatically allocate the computation to the best architecture or environment.

**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.

We are especially interested in strategies and infrastructure for scholars to find each other across the disciplines. Support for interdisciplinary (No-boundary) training, collaboration, and consultation is needed. Scholars of computing and analysis need to connect with the domain scientists to understand the more pressing problems that are in need of technical solutions, and the domain scientists need means to discover the most up to date technological developments that they need to efficiently and effectively solve their problems.

- Statewide online consulting that is available to students, teachers and scholars. Training and consulting should include software, hardware and algorithmic strategies for most efficient computations, and support to match of those in need with the best technical environments and computing products. Also needed is assistance for scholars and students to become knowledgeable about various analysis, storage, and computing environments so that they are able to determine when to seek assistance of a technical, computational or analytical expert collaborator or consultant or seek extra training.
- Attention to the ethical concerns around the use of data, including impact of big data analysis on our natural environment (the carbon footprint), but also the implications of the interpretations of data analysis on people, society and our democratic system.
- Assistance and training in modeling. For example, how to structure data files for efficiency, how to design a relational database for consistency and correctness under well-formed SQL

queries, or how to design an experiment to support causal inference. • Collaboration and teamwork is difficult for some scholars who were educated in domain silos. Some training in teamwork is necessary. While social scientists have documented that diverse teams of experts are usually most effective in solving the most vexing problems, we need these findings to be translated broadly to the community of scholars. This is especially important as we all realize that research funding is increasingly targeting diverse teams of experts. Note: By “diverse”, we mean gender, cultural, social economic diversity, and scholarly diversity. This includes different domain perspectives as well as theoretical and practical/applied expertise, etc. • While connections to national storages and platforms are available, broad support for regional or statewide resources would also be very helpful. Strengthening local environments to match national resources when for example, overflow of data or computation requires movement back and forth among different local, regional, and national resources. • A regional or national record of the data resources that have been helpful to certain domains or groups. For example, a catalog of best solutions/practices, software choices, strategies for efficient computations, etc.

-- End Submission --