NSF 20-015 Dear Colleague Letter: RFI Due Dec 16th

https://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf20015

NSF invites both individuals and groups of individuals to provide their inputs via the online submission form (link below). The submission form requires the following information[1]:

- Contact person name and affiliation. Lisa Arafune, Coalition for Academic Scientific Computation, CASC
- Valid contact email address. Lisa.Arafune@casc.org
- Additional author name(s) and affiliation(s).
  - Erik Deumens, University of Florida
  - Sharon Broude Geva, University of Michigan
  - Jonathon Anderson, University of Colorado at Boulder
  - Mike Warfe, Case Western Reserve University
  - Scott Yockel, Harvard University
  - Dhruva Chakravorty, Texas A&M University

- Research domain(s), discipline(s)/sub-discipline(s) of the author(s).
  - Advanced Research Computing
- Title of the response Coalition for Academic Scientific Computation (CASC) Response to NSF RFI 20-015

Three questions + abstract:

Abstract **(maximum 200 words) summarizing the response.**

The Coalition for Academic Scientific Computation (CASC), is an educational nonprofit 501(c)(3) organization with 93 member institutions representing many of the nation's most forward-thinking universities, national labs, and other research centers. CASC is dedicated to advocating for the use of the most advanced computing technology to accelerate scientific discovery for national competitiveness, global security, and economic success, as well as develop a diverse and well-prepared 21st-century workforce.

In this response, we have included some of the cross-cutting observations reflecting the needs and challenges encountered across our diverse membership. These needs and challenges are given through the lens of the support and enabling of data-intensive research, rather than that of specific research projects. Data management, defined broadly, is at the heart of many of these challenges, as is workforce development. Solutions for these challenges must come from closely interlinked initiatives, with funding and policy organizations working with the research and institutional strategic leadership communities.

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 power of data-driven science lies in linking large and diverse data sets. Bringing these data sets together for use in methods such as “machine learning” and “deep learning” requires metadata to properly connect data elements and/or standards in which the data are presented so that the meaning of data elements is known. This poses problems especially for interdisciplinary data where the potential gain in knowledge from connecting data is expected to be the largest. Solutions will require a consolidated effort across community boundaries to develop metadata best practices and standards.
- Tools and strategies need to evolve to enable the synthesis of information and data that is disciplinary agnostic. NSF/OAC/CISE has done much to create coherence between interagency technology and application portfolios. Are there opportunities to fund more investments in NSF’s CIF21 to further collaborations and innovation? Does NSF have enough resources to manage new opportunities effectively?
- Cross-disciplinary challenges abound where one major obstacle is the lack of interchangeable data formats. Some communities have created data standards, with the astronomy community as one example. A more precise guidance document and explicit requirements for creating the mandatory data management plan in any NSF proposal could help here. Such a document could specify and expand on required elements in the data management plan: the type of data format that will be used, what data will be shared at what stage during the project, and how data will be made available. Preferred standards, if these exist, should be noted in the requirements. Once standards for data are required and mature within research domains, it becomes possible, e.g., within the NSF BigData Hub program, to create data repositories that take multiple standards from multiple communities and produce harmonized cross-disciplinary data sets for interdisciplinary research purposes.
- An additional obstacle for both domain and cross-disciplinary research is the ability to find data, even in cases where institutional- or domain-repositories exist. Data that can not be found easily (along with provenance and other metadata) will be regenerated, recollected and rehoused, without the benefit of previous processing and cleaning of those data, or insights gained while manipulating them.

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?***

- In end-to-end scientific data-to-discovery, the path often involves large data flows created by sequencing, microscopes, telescopes, and spectroscopy instruments. This type of data flow still has a huge *first-mile* problem, which is due to the heterogeneity of deployments in experimental labs (e.g., ill-configured networks, lack of modern protocols). Since these instruments are outside of the data center, they are normally supported by local researchers or Enterprise IT, which both are ill-equipped to handle this type of data flow when it comes to network, data-transfer, and storage technology. Reference architectures with standardized data flows and trained CI professionals to facilitate in edge-computing deployments will allow optimizations to be realized.
- Diverse data sets are likely stored on diverse systems, some in the commercial cloud, some in community repositories, some in storage systems at academic systems. Protocols to access data efficiently without needing to move data more than necessary and without incurring large fees need to be developed.
- Increasingly, research will involve some data that is restricted and requires handling within certain compliance frameworks because of governing laws, regulations, or contractual requirements. By linking such restricted data with other data, the combined dataset, for the duration of its life during processing, will be restricted under the same requirements. Thus powerful HPC environments will be needed to work with such large combined data sets from diverse provenance.
- The end-to-end scientific workflow challenge is a deep problem that has been known since the early days of TeraGrid and persists during XSEDE: Most researchers pick a system and do all their work on that system. Multi-institutional research teams know this problem when using their campus resources mixed with commercial cloud resources, where each campus may have a different commercial cloud provider as preferred and approved provider. Several NSF funded projects are working to explore this issue, such as the Pacific Research Platform and the Aristotle Cloud Federation. A more systematic

and comprehensive approach may be needed to provide useful guidance for the community. Maybe this can be a topic for a CI Center of Excellence?

- As data services are developed, especially those that involve broad access or those that involve commercial storage and compute services, careful attention must be paid to intellectual property ownership, licensing, and compliance accountability.
- Current NSF data storage requirements are often in conflict with the data-storage service funding models available at research institutions. NSF should support and solicit the development of national data-storage CI that can be granted in-kind, similarly to the granting of compute cycles today.

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.*

- Data scientists and data engineers arguably represent some of the most sought-after workforce in the world. Two issues must be addressed: 1) The workforce lacks a common set of skills that form the basis of disciplinary-agnostic and data-agnostic collaborations. 2) Access to the talent required to solve data-intensive/data-driven S&E research challenges.
- The business models for data-driven science need to be considered. There are several questions:
    - Not all data is valuable. Metadata needs to be created with the data to indicate its value e.g. in terms of retention policy.
    - Data that is considered valuable must be considered as such by a number of stakeholders who show commitment to the value assessment by contributing to the cost of keeping the data.
    - The cost model of “pay-as-you-go” in the commercial cloud eliminates the option of scavenging for idle cycles, a model that has been demonstrated valuable for both learning- and high-risk-of-failure exploration of ideas. This mode of operation has been of value on research clusters for decades and is at the core of the Open Science Grid. This is just one aspect of where the research and education community needs to negotiate different terms if we are to continue to use commercial cloud providers.
- As data-intensive research becomes more and more common across disciplines, there needs to be a “bigger picture” view of workforce training and education. It is no longer sufficient to rely on separate training that provides basic coding and systems skills or higher-level domain knowledge but does not provide computational- and data-intensive training and curricular education that combines both.