

Reference ID: 11227067377_Stanzione

Reference ID: 11227067377_Stanzione

Submission Date and Time: 12/16/2019 4:53:03 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: Dan Stanzione - The University of Texas at Austin

Additional authors: Kelly Gaither, UT-Austin; Niall Gaffney, UT-Austin; Tommy Minyard UT-Austin; Maytal Dahan UT-Austin; Paul Navratil UT-Austin; Matthew Vaughn UT-Austin

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

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

Astronomy and astrophysics, cell and developmental biology, computer engineering, computer science, aerospace engineering.

Title of Response

A Center/Community perspective on Data infrastructure

Abstract

In this response, we attempt to summarize issues we have seen from a Center perspective across multiple disciplines, and in the settings of both large community projects and individual research

projects. We do not propose any particular “holistic” solution to these problems, but rather point out certain gaps in current infrastructure or areas where further investigation into the appropriate type of infrastructure may be warranted. While the response is meant to be cross-cutting, we are drawing these comments from experience with a number of communities: * The NHERI (Natural Hazards Engineering Research Infrastructure), a set of experimental facilities and community of researchers that deal with the design and response of the built environment to hazards. Parts of this community have been building shared data infrastructure for almost 20 years. * The Cyverse/iPlant community, an NSF-funded community infrastructure originally targeted at plant biology, now in its 12th year. * The Synthetic Discovery and Design (SD2) Community, a more recent DARPA program bringing together diverse teams of data providers, modelers and analysts, and infrastructure providers. As well as many individual users of NSF and state funded infrastructure at the Texas Advanced Computing Center, through the XSEDE project and other means.

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 broad category of “Data-Intensive Research” represents not one challenge, but many. Some of the questions are discipline-specific, and in many cases the actual research questions evolve iteratively as the data is analyzed and explored. But, there are a number of cross-cutting research issues that could benefit from shared infrastructure. Fundamentally, there are simply the questions of how we deal with the decreased cost and consequent increased volume of data to store, analyze, share, and potentially save. But beyond those challenges, there are many issues to deal with beyond volume. Many of these questions revolve around AI and Machine Learning. To list a few: How can surrogate models built through Machine Learning inform the simulation workflow? The performance of simulation science can be dramatically enhanced by using surrogate models to prune parameter spaces or even replace the complexity of computing physics-based models. This requires rich datasets (from experiment or simulation) to train the model on the physics, and extensive validation of outcomes. Trust in these models will come slowly, unless we can also answer the next question. . . How can we verify, and validate results given from Deep Learning/AI experiments, and verify that it is free from bias? Any model will not only need to be trained, but also will need methods to confirm the validity of results; in some cases, this may mean validating candidate solutions from AI with a full physics model; in others, it may mean adjusting the methodology to use an AI method where decision points and criteria can be clearly traced. There are many other questions relevant not only to AI but to data science more broadly, for instance, * How can we assess the quality of data we are using in our analysis? Repositories today offer a lot of data; but often little indicator of the quality. Was it collected by a Nobel prize winner or a high school intern? What is your metric of trust? * How can we assess whether data will be re-used? Broadly, we believe research data is valuable and try and retain it for reuse. Clearly, it costs a great deal of time and money to collect. But does the simple existence of this data mean it will be re-

used? What types of data lead to re-use? Are there things inherent in data collection or experimental design that promote or hinder re-use? These questions aren't limited to the data themselves, but how they are prepared for re-use as well, for instance: * Is the properly formatted for re-use, with sufficient metadata? How much is enough? Who should bear the cost of this? * Can we reuse unstructured data without human curation through AI-driven knowledge discovery? * How do we match data with analysis in a way to support reproducibility? A question that is not scientific, but certainly impedes progress, is whether there is a way to re-use existing datasets in research that has minimal friction and predictable costs, particularly when sources are varied or reside with commercial clouds.

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 first element of required data infrastructure, currently somewhat lacking, is, simply put, a place to store data. A place to store it for a long period of time, where it does not change, migrate, or disappear when needed. In essence, a data library. Many current attempts at solutions in this space focus on commercial solutions (e.g. through publishers), public clouds (pay-as-you-go access), or small focused repositories, limited in either discipline or regional scope, with attempts to federate these distributed systems together as a single whole. We would argue that these solutions are incomplete, and add needless complexity. Federation is a popular approach, because it allows a single “payer” (say, a single NSF program) to not take on the entire scope of the problem. However, from a science perspective, it is not a desirable state. Every minute spent working on protocols and implementations and standards to make federation work is overhead. The end state of federation is almost inevitably not seamless, nor sufficiently robust (if it were, simply giving all data a URL would suffice — the web is a lowest-common-denominator federation, but it's advent 30 years ago has not solved our data problem, or reduced the amount of required development). Large, distributed systems are hard to build and maintain, prone to faults, and any successful ones require a degree of uniformity that is difficult to enforce. Distributed software systems is a rich research area, but not necessarily a practical data solution. Infrastructure is also needed to address the disconnect between “repository” and “working space”. Too much focus is on simply the place to format and deposit the data, and not on how this data is consumed. It is critical that data be co-located with analysis capability (computing hardware, software tools, and expert people) and not assume that simply having the data present is sufficient. Interactive exploration is also key — rarely can a simple analysis be coded and run and answer the research questions posed by a

rich data set. Probing the complexity of data takes judgement and iteration; each query of a dataset leads to the next, and only interactive exploration gets the researcher to the ultimate answer (a related issue is how we visualize these results at scale — conventional scientific visualization techniques break down in the face of the volume and variety of data that we face). Interactive exploration and discovery needs to happen without surprise costs. Discovery of data in these environments need to deliver data in usable forms; long lists of URIs to web pages describing data is not sufficient for automated analysis. Interfaces need to support new users but also experts; every web interface needs matching API and CLI functions. Many data producers have limited tolerance for complex schema and ontologies or laborious manual curation requirements. The likely future of data discovery will be AI-generated metadata and categorization of largely unstructured data; there will never be sufficient budget for human-curated data given the tiny cost of producing large volumes. Communities around data are built over time — sustaining these communities and their practitioners will be essential to progress.

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.

Workforce remains a significant issue. Not simply the training of “data scientists”, but a community of researchers spanning all the necessary “data” skills, with some deeply-trained practitioners and AI and machine learning, statistics, visualization, computation, etc, but also broad community “literacy” in these topics. Workflows will be inherently more complex. A larger issue is around sustainability. Our incentives in this space are misaligned. We must consider who receives the *value* from making data sustainable, and how can our strategy be made incoherent. A common fallacy is that sustainability is incumbent on the cyberinfrastructure provider. However, computing and data skills are in high demand. It is far simpler to move on to the next “hot” well-funded project than to scrape together funding to sustain one. If employment isn’t the motivator, there are certainly few rewards — maintaining data sets will not lead to publications, promotions or other accolades for the maintainer. An incoherent sustainability strategy is also not sustainable. A “micropayment” model will not be effective — using tools that have multiple plans for sustainability is a problem. The value of data holdings is to the consumer of the data, and more broadly to the research community in paying less to generate the same data repeatedly, in turn freeing up funds for more research. Rather than making it incumbent on CI providers to find individual models to sustain data (when more attractive work is abundant), the community need methods to make sustaining datasets a *desirable* activity that attracts providers.

-- End Submission --