

Reference ID: 11226987378\_Hoban

---

**Reference ID:** 11226987378\_Hoban

**Submission Date and Time:** 12/16/2019 4:20:02 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: Sean Hoban - Center for Tree Science, The Morton Arboretum

**Additional authors:** Alissa Brown, Center for Tree Science, The Morton Arboretum Andria Dawson, Mount Royal University John Robinson, Michigan State University Adam Smith, Missouri Botanical Garden Allan Strand, College of Charleston

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

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

Population Genetics; Demographic Modeling; Conservation Biology; Environmental/ Ecological Modeling; Paleoecology; Organismal Biology; Invasion biology; Evolution; Community Ecology; Genomics

#### **Title of Response**

Computational challenges with accessing, merging and utilizing diverse and large datasets in ecology and evolution

#### **Abstract**

The scale of data requirements and availability is increasing rapidly, but infrastructure allowing researchers to freely access, manage and use robustly maintained and quality-controlled data is lacking. In ecology and evolution, we increasingly need to integrate different data to answer challenging questions that range from local to earth system-level scales. Furthermore, these data sources are accompanied by unique sources of uncertainty and varying resolutions, and some data come from models with their own set of assumptions and variables. Merging varied data sources represents a core challenge across science and engineering, where data availability has outpaced the creation of infrastructure and software needed to accommodate researchers with varying levels of expertise in data science. Here we establish some critical data science needs in science and engineering, including: data storage, management, validation, updating, version control, and accessibility, and the training required for researchers to use the data; linking disparate datasets, allowing researchers to avoid lengthy and error-prone data restructuring tasks; accessible cloud computing; canonical datasets; and standardizing analysis pipelines that are useful across computing platforms. Finally, we discuss the opportunities such requirements can afford science and engineering research, such as enhancing interdisciplinary work through synthesis centers and collaborative grants.

**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 scale of data needed to answer novel questions in science and engineering has been expanding, which necessitates new approaches to use dataset of varying size and complexity. In the field of ecology and evolution, a core challenge is integrating different data types, each with different attributes, modes of collecting, sources, resolutions and associated models. A major question in our field is: how have species, communities, ecosystems, and earth systems responded to environmental change over thousands to millions of years, and how will biodiversity respond to future environmental change? We can only answer this by combining past and present data, across spatial scales. For example, we must integrate data from modern ecology (occurrences, observations), paleoecology (fossils), genetics (genomes), and geography, and fairly quantify uncertainty stemming from each of these data sources. Other specific questions showcasing the need for cross-disciplinary data merging in the field of ecology and evolution include: To what degree have species and populations exchanged genes (and adaptations) over evolutionary time (thousands to millions of years), and how can we reconstruct phylogenies and demographic histories that accurately represent this history? How can we distinguish true genetic signals of adaptation using gene-environment association methods? What are the most important causal mechanisms of large-scale patterns in genetic and species diversity globally, including organismal traits, geographic range size, latitude, and local species richness? These questions all require methods for combining varied data sources, from different data repositories, which each have limitations in terms of data usability and validation. We believe that these challenges are faced by many disciplines in science and engineering, and can be used as an example for developing new ideas to enhance data

management practices, accessibility, usability, and robustness. Some of the principal challenges we will detail in the next question include: storage and accessibility of data (especially for large and complex data sources such as satellite images, NEON data, GenBank, and eDNA data), connecting datasets that share linkages (i.e. occur in the same geographic or temporal location or from the same species or community, as in the database Cartogratree), standardizing metadata, quantifying uncertainty in data, and developing standard analysis pipelines. Of course, access to and training in use of and best practices for high performance computing (HPC) is a need for all. Sustaining such resources will require funding on longer timescales than is typical for NSF (5 to 10 years or more).

**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 propose several advancements needed for data driven scientific discovery; these arise from the desire to connect datasets currently in distinct formats and databases not designed to work together. Next-generation ecology and evolution will require integration of versioned datasets from multiple sources; modular, flexible, open-source software; applications that work across computing platforms; and embedded quality-control enabled by development of canonical datasets. Of course, the software needs to be flexible, modular, and open-source, ideally with community driven documentation, periodic training, improvement and maintenance. Current online databases in biology differ with respect to data availability and quality. Many fall short on: ease of query, ability to batch upload and download, metadata, and version control. For example, most occurrence data has to be cleaned (time-consuming) or discarded. Few population genetic datasets are georeferenced requiring extensive work to assign locations to genotypes. Thus, one pressing need is a system allowing end-users to uplink corrections or additions, with accompanying metadata and version tracking for each datum. Another need is to record data validation from data collection onward, e.g. a record of a series of (standard) checks that data has or has not gone through. We are unaware of any publically-accessible collections database that allows this (e.g., BISON, GBIF, Neotoma, iDigBio, Genbank). Most databases also need to be more user friendly for flexible searching and downloading. Meanwhile, existing and new databases need increased amount of and standardization of meta-data (e.g. GenBank). Lastly, databases do not exist for key sources such as within population genetic allele frequencies, satellite images, soils, and eDNA. Moreover, digitization of historical data (such as from supplemental materials, unpublished data, grey lit data, theses) is needed, as are improved tools for citizen science projects. An ongoing challenge is the need to frequently switch between software and platforms to complete analysis. This entails time-consuming

and error-prone restructuring of data. Simultaneously, in the technology sector, computing is undergoing a re-conceptualization from a product to a service (e.g. from on-site to cloud-based systems). Thus, we see a need for cloud-based systems that: allows scientists conduct analysis by pulling data sets from across the internet using existing and new tools (e.g., R, Python, Julia, GIS), with tracking of versions of software, data, and (especially) analysis pipelines, and produce final output ready for publication. This would dramatically increase efficiency and reproducibility in science. Conceptually, such a system would combine the user-friendliness of a markdown document, jupyter notebook, GitHub, or Cyverse page with more advanced function. Also important are user-friendliness and flexibility. Cloud-based analytical pipelines should be approachable for students and researchers without computer science degrees yet powerful enough to accomodate advances in AI and machine learning. Recording analysis pipelines, including the decisions made by practitioners and the uncertainty that is propagated at each step, is crucial. As scientists become more removed from analytical details and perform more “big data analysis”, they are less likely to notice mistakes and dataset particularities. As a result, increasing computing access and power increases the risk of propagation of errors across hundreds or even thousands of publications that depend on a single analysis system. Thus, we propose building on the concept of “canonical” data sets like Fisher’s irises- we encourage the development of canonical “constellations” of datasets that should yield a known set of results when combined in cross-domain analysis. This will likely include simulated data. This will be critical for testing methods that are too complex for any one person to evaluate and become increasingly important as AI and other black-box methods become a standard tool across the sciences.

**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.

Cross-disciplinary collaboration to address scientific questions in twenty-first century biology requires deliberate coordination and leadership. In addition to training highly qualified personnel in traditional disciplines, it is essential to develop skills that allow for collaborative projects. These skills can be developed through a combination of online training modules, in-person workshops, and through activities at synthesis centers. Additionally, collaborative projects to address cross-disciplinary questions will require funding support from organizations like the National Science Foundation. Therefore, it is important to maintain and expand collaborative grant opportunities, in particular opportunities for international collaboration with researchers in Canada and Mexico because ecological and evolutionary processes cross national boundaries. More specifically, successful collaborative efforts require exemplary communication skills, and the ability to balance the needs and interests of both teams and individuals. In natural sciences, recent research addressing large-scale scientific questions also requires computational, data processing, and often quantitative modelling expertise. Access to not only training opportunities, but also resources that permit this type of research is required. With the increasing disparity among institutional access to cluster computing, it is necessary to take action to improve accessibility to these resources. Finally, it is important that integrative, collaborative projects assess the improvements made possible by expanding sample sizes, using advanced algorithms, and/or combining

datasets. The novel approaches that may be integrative and often use “big data” must be compared to and validated against established approaches and scientific understanding. This evaluation helps elucidate potential biases or pitfalls. Therefore, evaluation of newly developed software and methods, and comparisons to existing approaches, should be encouraged in funding opportunities, perhaps including funding calls focused on method evaluation and cross-comparison.

-- End Submission --