

Reference ID: 11227701661\_Sun

---

**Reference ID:** 11227701661\_Sun

**Submission Date and Time:** 12/16/2019 11:24:28 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: Ziheng Sun - George Mason University

**Additional authors:** None

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

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

Geospatial cyberinfrastructure; GIS; Remote Sensing; Agriculture; Machine learning.

#### **Title of Response**

Building Glutinous Infrastructure for Managing Disparate Data, Algorithms and Processing Services

#### **Abstract**

The cyberinfrastructure for science has undergone a rapid development phase in the past several decades. Scientists now can easily access a tremendous amount of observational datasets and find many software/services to analyze them. However, the existing of numerous datasets and platforms brings a new challenge to scientists to learn and use them, especially when multiple datasets and software are

required. One of the urgent needs in science is developing some glutinous systems through which scientists can manage their access to distributed datasets in one place and use the disparate services/software to process the datasets.

**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.

My main research field is agricultural remote sensing which is a disciplinary research integrating GIS, RS, big data, machine learning, sensors, biology, geography, and knowledge from many other domains. Recently the most popular topic is using machine learning to train models on historical datasets and predict the real-time/future maps of crop distribution, status, and yield. There are three major challenges in achieving that goal: big data fusion, lack of data labels, and efficient learning on big training data. Fusion of the big disparate data is a long-lasting problem in the field. Satellite remote sensing data has continuous large-scale observation of the crop fields but full of clouds and noises caused by other meteorological events. Fusing the observations from different satellites like Landsat and Sentinel could remove a few but not all of the gaps. Airborne images (drone) or ground measurements are more accurate, free of clouds, and have less noises. Data labels are essential to train the existing machine learning models. Normally the labels are manually annotated from the raw observations and take a lot of investments on labor and time. The third challenge is the inefficient learning on big data. Scientists have to spend a lot of time on initializing experiment environments, installing required software libraries, and writing scripts to get the model training/testing running. As remote sensing data is normally stored in data centers with many nodes, scientists have to deal with more than one machines/platforms (clouds) in their experiments. Efficiently managing all the resources and running models on top of them are very challenging.

**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 agricultural remote sensing, we already have cyberinfrastructure like Google Earth Engine to help us quickly load and analyze remote sensing data based on uploaded sample datasets. However, Google Earth Engine (GEE) doesn't directly support the training of many machine learning models like the deep neural networks which are very popular in these days. We need use other systems like Colab to get models trained and then import the trained model in GEE for prediction. For most scientists, they are exhausted to study and get themselves to fit in these new tools. They need a one-stop tool which can allow them to build, create, run, monitor, and review the entire machine learning workflow in one place. They want to be familiar with one entry tool which offer a hub of the capability of the existing popular tools on machine learning and the tool doesn't require them to learn too much knowledge and get used to brand new interfaces.

**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.

All that matters in cyberinfrastructure is people. The building, running, using, maintaining, and retiring of web systems should focus on the urgent needs of its associated people. In agricultural research, the end users are farmers, crop stakeholders, agriculture departments, insurance companies, food consumers, industrial users, etc. The proposed design of cyberinfrastructure should take user requirements into account since the very beginning, e.g., including user representative in the team would have very profound benefits. The developers of cyberinfrastructure could learn from the real users and significantly improve the success rate of the projects.

-- End Submission --