

Reference ID: 11089184599\_Just

---

**Reference ID:** 11089184599\_Just

**Submission Date and Time:** 10/23/2019 10:57:32 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: John Just - Iowa State University

**Additional authors:** None

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

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

data science

#### **Title of Response**

the "IO" bottleneck

#### **Abstract**

With GPUs readily available that can calculate gradients reasonably fast, the main bottleneck in training AI models is the "IO" speed. Most big data is accessed over a network, not locally. Training a model from data on a local SSD and accessible via PCIe is many times faster than the same over a network.

**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.

We are collecting huge amounts of data and have been innovating in the AI space for commercial applications, but getting data pre-processed and to the GPU efficiently can be a daunting task (pipeline optimization) and sometimes not possible at reasonable costs to many researchers or companies. Data stored on SSDs and accessible over a PCIe bus is many times faster for training a model than trying to do it on "the cloud". This really limits the progress that can be made and ends up with us dumping a lot of time into trying to optimize IO pipelines and data formats.

**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?

Again -- there is a lot of data out there but the speed of access is painfully slow when stored on the cloud....severely limiting progress on model training and exploring data. But then it's extremely inefficient to have to store data on drives physically connected to the machine with the training hardware, or sometimes not realistically feasible at all since most data stored on the cloud can be much larger than local available storage. Prefetching data with parallel input pipelines is not the solution and does not overcome IO/network delays.

**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.

-- End Submission --