## Introduction

This response to the NITRD RFI and draft Big Data R&D vision is submitted on behalf of the authors of a workshop report on data science priorities in the Earth and environmental sciences. In the report, the participants identified the need for comprehensive, cross-agency efforts to address existing and looming challenges related to the creation, management, use and re-use of agency sponsored scientific data. The participants strongly urge the following initiatives be put in motion related to science data challenges: 1) enable the National Research Council to lead a high-level task force study of these issues to provide strategic guidance that agencies can use in planning, 2) consider the creation of a cross-agency effort to create a sustained Science Data Infrastructure (SDI). The full document, *Workshop Report: Planning for a Community Study of Scientific Data Study,* including an executive summary may be found at http://dx.doi.org/10.7269/P3R49NQZ. The workshop was funded by the Gordon and Betty Moore Foundation, by the National Consortium for Data Science (http://data2discovery.org), and by ESIP (Federation for Earth Science Information Partners, http://www.esipfed.org).

## Genesis of the Workshop / Workshop Participants

The data science workshop arose from within the Earth science community as represented by ESIP. ESIP and its members epitomize the broad swath of Earth science communities and agencies charged with observing and monitoring the Earth, including NASA, NOAA, USGS, and EPA. ESIP formed a working group to consider the idea of a data science study and ultimately led the way in defining, organizing and leading the workshop. The participants included recognized leaders in the fields of Earth and environmental science and data science. While the workshop focused on Earth science and related domains, participants recognized that these challenges and findings were likely pervasive across the spectrum of scientific and applied research.

## Summary of Needs Identified

The report identifies five categories of challenges around scientific data: 1) economic, 2) cultural, 3) data science research, 4) educational, and 5) legal, ethical, and policy.

The primary challenge relates to the economics of data. We do not know the value our data provide. Though, research is growing into how to measure both actualized and generative value of research data. Early results show significant positive returns on investment from two to twenty times the cost. Currently, data management costs are typically scraped from research funds; with the associated tension and resistance that situation creates. To create an effective, sustainable funding model the economics need to be better understood.

Cultural changes are needed to recognize and reward the value of dataset creation, management and stewardship. Data science research includes questions about data discoverability, access, interoperability, stewardship, and a possible mathematics or science of data, among others. Data stewardship must be integrated into science education, and appropriate data must be accessible for science education at all levels. Finally, legal, ethical, and policy challenges include licensing issues, achieving scientific goals of verifiability and repeatability, as well as the possible creation of an agency, council, or

office to support scientific data infrastructure.

***Recommendation #1: Convene a high level, fast track study of cross-agency needs and future strategies. The study should be conducted by the National Research Council as a highly respected neutral party.***

The participants in the workshop strongly urge the creation of a high-level National Research Council-led study "to conduct an authoritative, unbiased assessment of strategic scientific data investments." The purpose of the assessment, as described in the report, is to provide strategic guidance on improving the management of scientific data in a cost effective manner to support the national interest. In addition, the study is intended to establish or re-establish the United States as a leader in science data infrastructure. Participants also encourage the use of an approach to the study that emulates a high profile panel approach. A long, drawn out traditional study is not warranted in this case. Any proposed study would also need to review and synthesize relevant prior work in science data management and infrastructure.

***Recommendation # 2: Envisioning a National Science Data Infrastructure (SDI)***

Workshop participants had a bold vision for a sustained Science Data Infrastructure (SDI) and associated technical and cultural shifts to better enable science in the face of major challenges now and into the future. In order to maximize the return on government investments and to minimize the duplication of systems and efforts, a cross-agency entity—something more than a working group or task force-- could be charged with attacking this problem. The end uses are fundamentally cross-agency, cross-disciplinary, and across temporal and spatial scales. The workshop participants recognized that science questions are increasingly complex and require inter- and transdisciplinary approaches for conducting research. The envisioned SDI would support science by providing a common framework under which all science data would be governed. The challenge is to transcend traditional agency missions in favor of a bigger, grander vision.

## Conclusion

The workshop participants are well placed to assess the types of challenges facing US Federal agencies in regard to data. These individuals work with the agencies and the data they generate, directly or through sponsored activities, in an ongoing basis. These individuals also work with various groups such as BRDI, CODATA, ESIP, RDA and others to exercise leadership in these areas. We look forward to engaging with NITRD in this effort as they consider these recommendations and the others submitted in response to this RFI.

## Point of Contact:

W. Christopher Lenhardt, Domain Scientist, Environmental Data Science and Systems Renaissance Computing Institute, Chapel Hill, NC, clenhardt@renci.org, 919-445-0480.