Author identification as requested by RFI. Brian Wee, Ph.D. || Chief of Strategic Alliances || National Ecological Observatory Network (NEON), Inc. || Email: bwee@neoninc.org || Phone: 202-370-7891 || Ph.D. (Ecology, Evolution, and Behavior), M.Sc. (Computer Science – Artificial Intelligence), B.Sc. (Information Systems and Computer Science) || Board Director on the Foundation for Earth Science that supports the Federation of Earth Science Information Partners (ESIP) || Member of the US Delegation to the Belmont Forum E-Infrastructure and Data Management Knowledge Initiative.

Response to RFI question: From an interagency perspective, what do you think are the most high impact ideas at the frontiers of big data research and development?

Environmental Data, Information, and Knowledge (EDIK) in the context of the three "V"s. The BDSSG appropriately included "diverse" in the draft vision statement, which characterizes the bulk of highly complex environmental data. It is also associated with one of the three "V"s of big data: volume (of data), velocity (of rate of data generation), and variety. While some datasets are characterized by the "volume" and "velocity" aspects (e.g. particle physics data), "variety" is particularly apt for describing environmental data that tend to highly heterogeneous. Within the context of this document, "environmental" includes the biosphere, socialsphere, atmosphere, lithosphere, cryosphere, and hydrosphere[4]. The Internet of Things adds an additional layer of complexity to the environmental data already observed by the nation's portfolio of civil earth observation assets[1]. The heterogeneity challenge is exacerbated by the complexity of coupled human-natural systems, in which any one variable (e.g. crop yield for any given year) is affected by multiple variables that each dominate at various spatial and temporal scales (e.g. soil bacterial processes subjected to diurnal cycles that ultimately affect atmospheric dynamics over a region over longer time scales) and demonstrate complex feedback loops (e.g. atmospheric processes that in turn affecting soil bacteria). Evolving the technology to enable repurposable, discoverable, accessible, and usable EDIK is crucial as society deals with the impacts of large-scale environmental change.

The case for evolving Big Data techniques for EDIK. The need for society to craft responses commensurate with likely significant perturbations to the food, energy, and water nexus has never been more urgent. Forecasting and managing such perturbations across both short and long-term temporal and geographical scales remains a challenge, but is important because: (1) longer-term and spatially-extent environmental forecasting are essential for guiding national and international policies for climate mitigation and adaptation, while (2) shorter-term and region- or sector-specific forecasts are essential for actionable environmental intelligence. This clearly requires increased interoperability between agency systems that are relevant to EDIK.

Hybridized data discovery techniques. There should be continued support for R&D to develop hybridized data discovery techniques that combine (1) statistically-based analytics and (2) heuristic- and logic-based semantic and ontology approaches. The inherent complexity of EDIK often breaks algorithmic approaches because of the myriad ways that environmental data are collected, described, and tagged (but are amenable to statistical approaches that establish numerical associations between patterns in data), while statistical approaches often lack the explanatory power offered by algorithmic approaches for humans to apply the "reasonability test" to the rationale behind machine-derived solutions.

Reference architecture for collecting, processing, and serving heterogeneous environmental data. Standards organizations such as the Open Geospatial Consortium (OGC) and the World Meteorological Organization endorse community-evolved conceptual frameworks (e.g. OGC Observations and Measurements model) and technical standards for representing environmental data (e.g. OGC WaterML 2.0), although most of these have been more extensively applied to natural (as opposed to social) phenomena. A reference architecture is needed that encapsulates these frameworks and standards into an end-to-end framework (a "multivariate end-to-end pipeline" of sorts) that starts with data ingest from in-situ sources to remotely-sensed instruments, organizes these high-dimensional environmental data in a scalable fashion that is amenable to hybridized data discovery techniques (see above), and exposes both data and processing services to authorized users via web services. The latter is especially crucial to enable third-party tools like cloud-based collaboratories. Such a reference architecture will ameliorate the risk of agencies implementing solutions that are not interoperable with others.

Cloud-based collaboratories. A cloud-based collaborative workflow tool may be thought of as a visual programming environment that a multi-disciplinary team would use to work on environmental challenges that ingest and produce complex EDIK. Workflow tools not only enable reproducible science, but improve the accountability of decisions by providing traceability for policy recommendations back to science and data. Workflow provenance conformant with the World Wide Web Consortium's PROV standard will allow those workflows to be harvested by third party visualization tools for visualization, re-use, and mash-ups. The DARPA funded PLINY project is developing technology that would offer the ability to "auto-complete" code by matching the implied objectives of the developer to the code-base in publicly accessible code repositories (GitHub, SourceForge, SVN, etc). A similar initiative for EDIK workflows (which can be thought of as pseudo-code in some ways) would reap big benefits in terms of accelerating discoveries by capitalizing on existing workflows and code. Big data also calls for compute resources to be located in the "vicinity" of voluminous datasets. In the context of workflow tools, this calls for further R&D into distributed processing and authentication and authorization infrastructure (AAI) that allow workflows to initiate jobs on remote compute resources collocated with data. Some of these ideas have been explored in recent years at the Earth Science Information Partners (ESIP) meetings under the label of the "Earth Science Collaboratory"[2].

Capacity building. Similar to what was previously pointed out by NEON, Inc. to an earlier NSTC RFI in response to a question about data management technologies, we need to support capacity building for scientists and professionals so that "they (1) have the knowledge to adjudicate on resources that should be expended on such technologies within a given agency, (2) have the knowledge to interact with other agency personnel to foster greater interoperability and coordination between agencies"[3].

Reference. [1] National Plan for Civil Earth Observations, 2014, NSTC. [2] Dr. Christopher Lynnes, NASA Goddard Space Flight Center. [3] NEON, Inc.'s Response (http://goo.gl/bFCrmE) to the NSTC RFI for [1] (Federal Register Doc. 2013-26890). [4] Invited presentation before the PCAST, September 11 2013 (http://goo.gl/DTJgCj).