EarthCube Strategic Vision:
Data-Enabled Transformation of Geosciences

VERSION 0.9: JANUARY 10, 2016

EarthCube's vision is to transform the conduct of geoscience in order to catalyze scientific breakthroughs, accelerate scientific discoveries, and advance systematic knowledge for an improved understanding of the Earth's environment. EarthCube's ultimate goal is to catapult the science vision of understanding and predicting a complex and evolving Earth system and help address critical societal needs, from sustainable utilization of water, energy, and other resources, to effective preparation and response to weather patterns, sea level and climate. EarthCube will achieve this goal by fostering modern modes of scientific inquiry that infuse emerging practices and technological innovations in digital scholarship, data science and analytics, data and software stewardship, and open science, into basic research in geosciences.

EarthCube represents an integrated and integrative approach to advancing both the NSF Geosciences Directorate strategic framework, *Dynamic Earth: GEO Imperatives & Frontiers 2015–2020*, and the NSF cyberinfrastructure vision[1] by encompassing the necessary range of technical, social, organizational, and cultural aspects of cyberinfrastructure development.

Science Imperatives and Frontiers

The primary scientific goal of EarthCube is to enable significant progress in our understanding of fundamental Earth system processes, and to improve our ability to mitigate complex, large scale environmental problems, by fostering innovations in data science and information technology.[2] Accelerated scientific progress in such topics necessitates nested and data-intensive domain-specific studies at ever increasing spatial and temporal resolution, as well as multidisciplinary interactions across different science domains. Future major advances will depend upon knowledge discovery within science domains and at their intersections, improved information flow, and advanced data analysis and modeling tools, all of which are catalyzed by EarthCube.

The motivation for EarthCube-enabled science can be distilled into three essential science *frontiers*:

1. **To characterize** the key processes, interactions, causations, and feedbacks operating at and across different temporal and spatial scales within physical, chemical, and biological domains.
2. **To quantify** limits of prediction and better understand the constraints on and limits of data and model accuracy and utility.
3. **To deliver** a holistic, quantitative representation of critical physical, chemical, and biological states and fluxes, in order to inform fundamental science and societal decisions.

---

[1] Cyberinfrastructure Vision for 21st Century Discovery, NSF Report 2007
[2] Examples include: geospace variability and extreme events; planetary-scale (global) changes in the Earth system; human impacts of geohazards; and water and energy sustainability.

The *imperatives* for advancing these geoscience frontiers include enhancing capabilities to make significant progress in understanding, characterizing, and mitigating complex, large scale environmental problems, such as:

1. Understanding the consequences, impacts, and effects of planetary-scale (global) variability and changes in Earth systems, including recognizing the geophysical signal within the natural variability.
2. Understanding geohazards and extreme events, through the effective characterization and communication of uncertainty and relative risk.
3. Providing sustainable solutions for water, energy, and mineral resource use by defining mass and energy balances associated with past and present conditions to accurately project future states.

Addressing these frontiers and imperatives necessitates the development of sophisticated and rigorous data-enabled scientific inference methods, i.e., technologies for learning from complex and ever-increasing data volumes, as well as methods that will guide what/how data should be obtained for effective scientific analysis and understanding. Geoscience discoveries are increasingly taking place across traditional disciplinary boundaries but currently they are hampered by inadequate availability of data across science domains and tools to process, integrate, synthesize and share those data. Scientists need improved methodologies and technologies to optimize knowledge transfer and discoveries at these interfaces. EarthCube aims to enable transformative research by not only: **1)** helping to address the computational, data, software, and knowledge management needs of domain geoscientists following basic, curiosity-driven, and exploratory lines of inquiry; but also **2)** creating opportunities and mechanisms that will allow domain geoscientists and computational and data scientists to work together to solve complex problems that intersect multiple geoscience domains.

Technological Imperatives and Frontiers

In order to foster cross-disciplinary research that considers holistic models of geoscience processes at different scales, EarthCube will have an integrative approach on the technology side. It is both an imperative and a major challenge for EarthCube to leverage the extensive existing infrastructure resources as the foundation for an architecture that enhances their interoperation, accelerates growth to incorporate new advanced capabilities, and facilitates sustainability.

EarthCube is emerging at a pivotal time in the evolution of the geosciences. Scientific research products, whether data or software or articles, are not only of great interest to scientists in other fields but have well recognized societal value[3]. Open science practices are crucial to the dissemination of scientific knowledge, including digital scholarship, data and software stewardship, and routine reproducibility[4].

The advent of the digital era has opened previously unimaginable possibilities, bringing cyberinfrastructure resources to many scientific communities, bringing data-driven scientific research to new levels. However, many of these powerful technologies have had uneven adoption and are often not well integrated into scientific practice in many communities within the geosciences. A

---

[3] https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf[4] https://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf15052

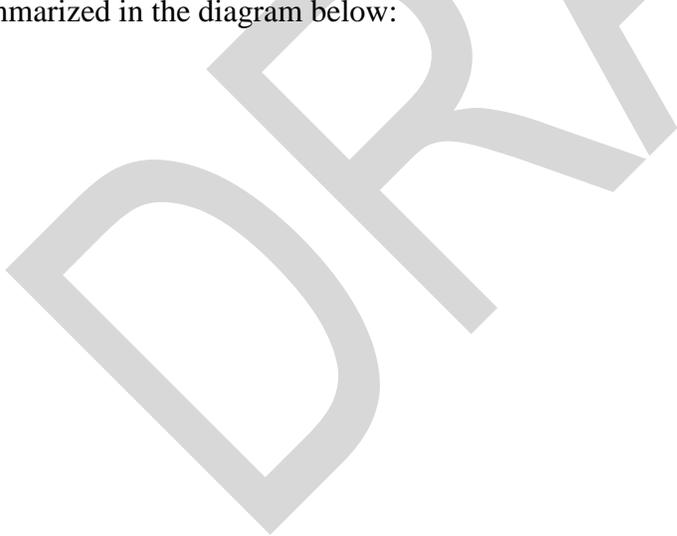[4] https://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf15052

more fundamental change is needed to foster active participation by scientists in driving the cyberinfrastructure requirements and providing continuous feedback to information technology professionals as tools and services are developed. Crucially, EarthCube cyberinfrastructure will be socially organic and driven by geoscientists' desire to discover and utilize multidisciplinary and multiscale datasets from distributed repositories.

Therefore, the core tenets of the EarthCube technology frontiers are:

1. **Knowledge-rich environments**: Scientific resources and products will be described with rich metadata to enable understanding and reuse. Advanced analysis and visualization techniques will be developed to enable automated learning techniques from data to take advantage of knowledge-rich components for effective mining, fusion, and assimilation of data into sophisticated inference models.
2. **Open science**: EarthCube will promote the design and creation of discoverable and accessible scientific resources and products, to enable reproducibility, and ensure that they can be adapted to solve new problems.
3. **Federated organization**: EarthCube participants, from organizations to individuals, will contribute resources designed to interoperate through agreed standards and protocols. Rather than centralized control, EarthCube will provide coordination by fostering standards and integration.

The imperatives for advancing these technology frontiers include advancements through supporting research in all stages of data life cycle, namely, data collection, storage, and management; data analytics and inference; and data sharing and collaboration. The components of these imperatives are summarized in the diagram below:

Data Sharing and Collaboration
- Tools for distant data sharing, real-time visualization, and software reuse of complex datasets
- Cross disciplinary information and knowledge sharing; interoperability
- Remote operation and real-time access to distributed data

Collection, Storage and Management
- Data representation, storage and retrieval
- Data management policies, including access and dark data
- Communication and storage technologies with extreme capacities

Data Life Cycle

Data Analytics
- Computational, mathematical, statistical and algorithmic techniques for modeling high dimensional data
- Learning, inference, prediction and knowledge discovery for large volume and dynamic data
- Data mining to enable automated hypotheses, event correlation and anomaly detection
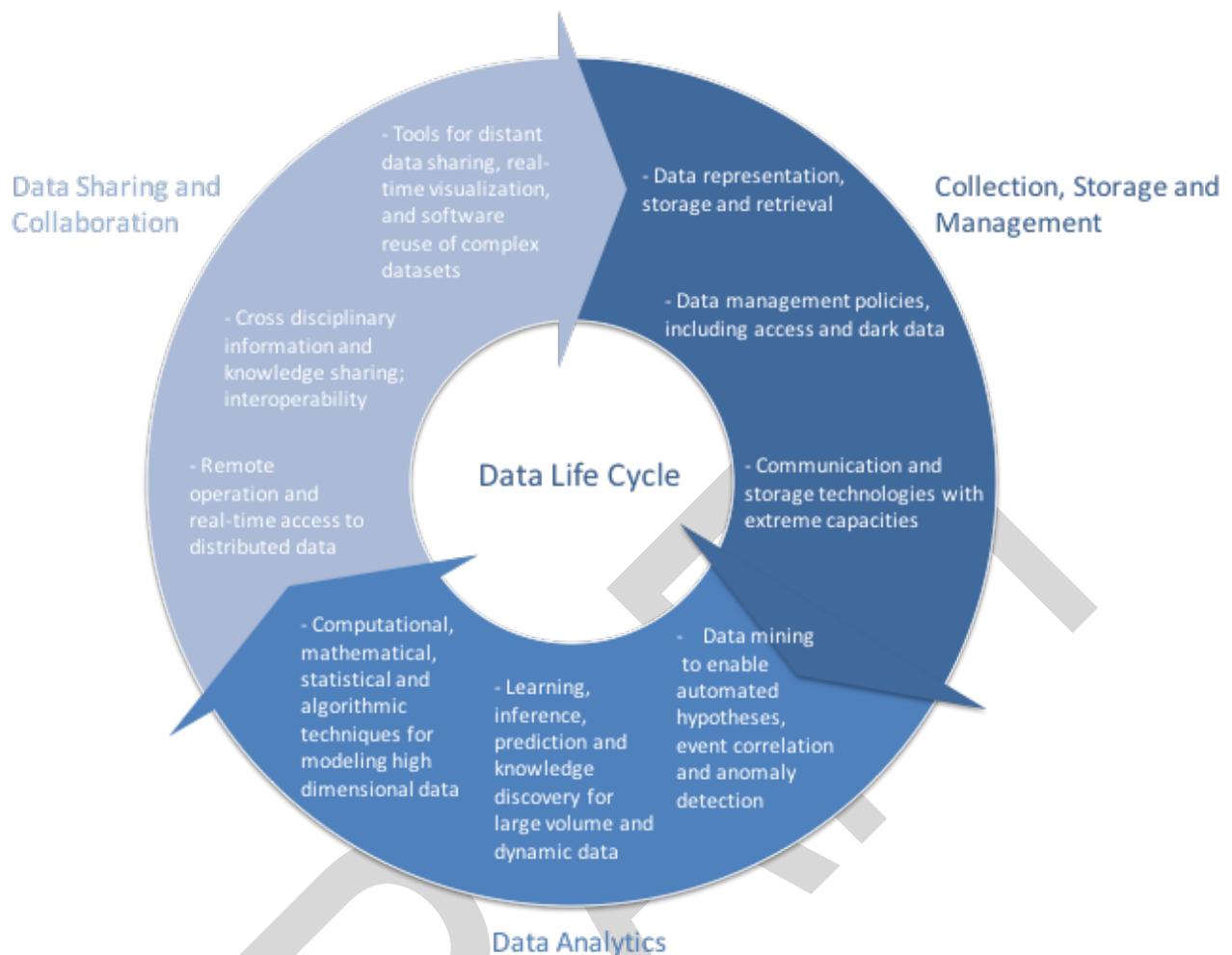
Figure Caption/Description:

Collection, Storage, and Management: Stewardship of contributed data (including rescue of dark data), software, and research products through open science practices will provide the foundation for all EarthCube-enabled science by providing access to high quality, interdisciplinary data necessary to characterize key earth processes across relevant spatial and temporal scales.

Data analytics: Knowledge-rich components will facilitate data mining and advanced data analysis, through computational, statistical and algorithmic techniques in order to quantify limits of scientific prediction, and constraints of modeling efforts.

Data Sharing and collaboration: Federated organization through the use of standard services will enable remote access, visualization and interoperability of complex distributed resources, allowing cross-disciplinary knowledge and information sharing to inform new science and societal decisions.

Recognizing that technology is the facilitator but not the agent for social change, EarthCube will endeavor to engage the community in adopting these principles. EarthCube will actively disseminate among scientists novel technologies designed to improve science practice. EarthCube will develop and promote best practices for interoperability and standards.

The degree to which different geosciences communities have been moving forward towards this vision varies greatly, as does the availability of cornerstone data repositories and the use of advanced cyberinfrastructure. Regardless of the democratization challenges, the overarching need in the geosciences is that repositories, tools and services are not linked to the data needed by scientists to address the contemporary science challenges that are identified above. Another common impediment is that in many cases data, tools, and other needed resources are hard to discover, access, and use. In order to realize EarthCube's science vision, geosciences disciplines will need to work together to overcome those barriers and move towards a more integrated environment for enabling multidisciplinary science and synthesis.

Capacity-building: Education, Training, and Workforce Development

EarthCube's goal of transforming the conduct of geoscience research is predicated on a vision for the fundamental role of education and training of the next generation of geoscientists in modern data-intensive modes of scientific investigation. Recognizing this intrinsically intertwined nature of scientific pursuit and human capacity building, EarthCube affords myriad opportunities to enrich the intellectual pursuit of geosciences by advancing data and information science literacy of students and early-career scientists. A highly-skilled geosciences workforce with strong abilities to effectively utilize research data and cyberinfrastructure is supported by broad-based capacity-building that includes 1) close collaboration with data and information science experts; 2) cross-disciplinary education and curriculum development as an integral part of research programs; 3) support for development of career paths and a cadre of technical experts.

There is an increasingly pronounced need for geoscientists to take advantage of rapidly developing and evolving cyberinfrastructure and data-related technologies. They need both enhanced skillsets in the realm of data science and close collaboration with a cadre of skilled technical experts. There is a compelling need to consider how to develop relevant skills in geosciences community in order to accelerate modern data-enabled modes of scientific inquiry. This requires meaningful and sustained interaction among geoscientists, data and information scientists, and educators to develop necessary skills in advanced data intensive scientific pursuit.

Among EarthCube's education goals is to foster activities to support systematic strategies for the development of education and training pursuit in data-intensive geosciences. In harmony with the commitment to interdisciplinary educational activities, EarthCube will also foster activities toward enhancing and refining pedagogy that infuses data-enabled research into geosciences.

Permanent URL:
http://earthcube.org/document/2015/earthcube-strategic-science-plan