Cyberinfrastructure Framework for 21st Century Science and Engineering

A Vision and Strategy for Data in Science, Engineering, and Education

Data are critical to discovery and decision-making in science, engineering, education, government, and commerce. The decreasing costs of acquiring data, the increasing density and complexity of data, and the increasing performance demands on data computation are transforming all areas of NSF-supported research[1]. Advancing science and engineering through data is a key element in the Cyberinfrastructure Framework for 21st Century (CIF21) Science and Engineering, a critical component of OneNSF, which is a concept that promotes collaboration in well-integrated and efficient ways across organizational and disciplinary boundaries.

The recent task force report of the Advisory Committee on Cyber Infrastructure[2] identifies several key areas in data and data management including: infrastructure delivery; cultural and social change; roles and responsibilities; economic value and sustainability; data management guidelines; and ethics and intellectual property. The NSB Task Force on Data Policies has put a priority on the development of effective data management policy and on ensuring the reproducibility of scientific results. These reports outline new challenges including quality control, analysis, documentation, provenance, confidentiality, security, sharing, and usability while at the same time noting the tremendous potential for the use, reuse, and repurposing of scientific data for entirely new discoveries and the urgent need for informatics expertise across all the disciplines.

This Strategic Vision identifies priority areas for NSF investment that will facilitate important and tangible progress in moving 21st Century science, engineering, and education toward effective use of digital data. The NSF vision is to provide a National Data Infrastructure that is advanced, robust, interoperable, scalable, and sustainable. This vision promotes greater balance in priorities, coordination, and leveraging, and encourages new strategies for maintaining prior cyberinfrastructure investments and creating new funding opportunities.

Strategic directions

The strategies outlined in this document contribute to the larger NSF CIF21 framework and are not separate or standalone efforts (see Figure 1). Although this document focuses on the data-specific strategies, it is important to note that CIF21 planning involves an integrative approach to

[1] PCAST Report: Designing a Digital Future: Federally Funded Research and Development in Networking and Information Technology, December 2010

[2] NSF Advisory Committee for Cyberinfrastructure Task Force on Data and Visualization. Final Report. March 2011. [PDF] Available from: http://www.nsf.gov/od/oci/taskforces/TaskForceReport_Data.pdf

support complex problems and issues addressed by science, engineering, and education communities. Implementation of the goals and strategies for data complements and dovetails with other CIF21 components, including advanced computing, software, learning and workforce development, and cybersecurity, as well as with individual directorate and office research and education efforts. Computational and Data-enabled Science and Engineering (CDS&E) and grand challenge communities' activities connect with the NSF National Data Infrastructure and all components of the CIF21 strategy. They are driven and enabled by a coherent approach to developing these components to meet the research and science requirements of the nation.

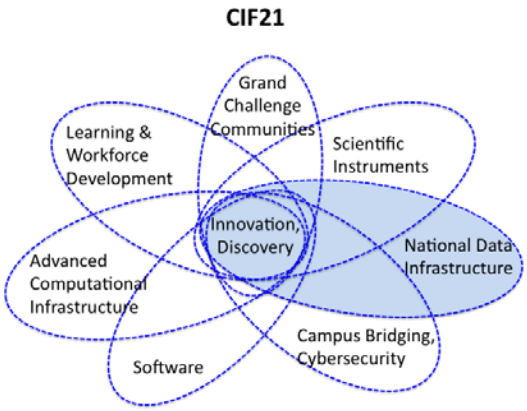

Figure 1

Goals for Data in Science and Engineering

In order to meet the challenges before it, the NSF will adopt, as part of its larger CIF21 mission, strategic goals for enhancing the role of data in science and engineering. There are five fundamental goals:

1. Provide science, engineering, and education with a national data infrastructure that will enable the capture, management, curation, analysis, interpretation, archiving, and sharing of data at unprecedented scale and complexity in a manner that will stimulate discovery in all areas of inquiry and from all facilities, ranging from the campus (midscale) to the national level (large facility) investments.
2. Ensure that this infrastructure stays at the most advanced state of sophistication, capacity, and capability through foundational research that both advances the leading edge of technology and contributes to fundamental knowledge.
3. Support transformative interdisciplinary and collaborative research in areas of inquiry stimulated by data through the development of robust, shared resources and the means for enabling partnerships across diverse communities.
4. Ensure that the future workforce of scientists, engineers, and educators are equipped with skills to make use of, and build upon, the next generation of data infrastructure.
5. Develop comprehensive policy for data, software, publications, and other digital outputs resulting from NSF-supported work. Such policy must acknowledge the new challenges surrounding reproducibility, storage, curation, and open dissemination of scientific data in all its forms, and recognize its importance for accelerating fundamental discovery, interdisciplinary research, and innovation in society.

The following sections detail the strategies envisioned for achieving these broad goals.

1. Developing and Providing a National Data Infrastructure

A national data infrastructure will have three essential functions, each requiring a directed strategy: (1) meeting the increased pace of data creation from advances in observing, experimental, and modeling programs; (2) ensuring the long-term sustainability, availability, security, and integrity of digital data; and (3) providing the tools, models, and algorithms to help drive discovery. These functions are connected in a recurring life cycle of data, as research generates new data, old data are revisited, and new theories are advanced. Each element must be usable, scalable, interoperable, sustainable, and available to support the current and future demand placed upon it.

1.1 Data Generation and Acquisition

Data are being generated at increasing rates and densities from a growing array of sources, including major surveys, mobile and embedded systems, sensors, observing systems, scientific instruments, publications, experiments, simulations, evaluations, and analyses. NSF will invest in technologies, methods, and services designed to facilitate this phase of the data lifecycle, by:

- Stimulating the development of a cyberinfrastructure that is closely coupled to the development of the instrumentation (from midscale to large facility) that generates data in real-time;
- Supporting research to maximize the benefit of high throughput observing systems by addressing both computational limits and workflow design;
- Develop scalable and flexible solutions for aggregating data from intrinsically diverse and disparate sources such as individual investigator research, crowd-sourced observations, and ad-hoc observing networks.
- Developing rigorous, requirements-driven designs and standards for data capture systems that are created and driven by end-users in science and engineering domains to insure derived data products and the workflows are well defined.

1.2 Data Storage, Curation, and Management

At the simplest level, a scientific data infrastructure consists of physical resources: hardware, software, and communications systems. A data infrastructure should (1) provide for the storage, persistence, security, and accessibility of scientific data; (2) have the capacity to accommodate the scale and heterogeneity of scientific data through robust, open, and broadly accepted standards; and (3) have an underlying, sustainable cost model that is fairly distributed across governmental, academic, nonprofit, public, and commercial stakeholders.

This infrastructure must also have the capacity for managing content, making it accessible and usable across a broad range of users from the providers to scientists, engineers, educators, policy-makers, managers, businesses, and the public. The range of scientific data to be

accommodated by this infrastructure varies widely from large, homogenous datasets to heterogeneous derived data products that require both maintenance and curation.

NSF anticipates making investments in data infrastructure targeting these areas:

- Scalable and sustainable physical infrastructure to accommodate current and projected data storage and communication and computation needs;
- Scalable and sustainable curation solutions to serve both domain-specific and emerging interdisciplinary science and engineering;
- Policy solutions that ensure sustainable operations and scalable security, access, confidentiality, and integrity to data;
- Partnerships with private, federal, and other organizations that share the burden of long term data stewardship across a diversity of funding models.
- Solutions to stimulate changing behavior with respect to data management and sharing practices.

1.3 Analysis, Modeling and Visualization

Science and engineering require more than merely the availability of data and associated tools. Analysis, computation, modeling and visualization techniques must be developed in close partnership with the domain sciences. NSF will invest in programs focusing expertise and resources to mobilize data, computational infrastructure, and software tools at unprecedented scales and capacities:

- Support development of a network of tools, algorithms and services, including hubs, portals, and gateways to provide national access and use of computer and data resources for analysis, modeling, and visualization.
- Address interoperability between distributed data sources representing a wide range of domains, data types, formats, and scales across science, engineering, and education.
- Encourage multi-disciplinary initiatives to stimulate development of tools and services in priority areas, and the extension and reuse of existing tools in other domains and research.

2. Foundational Research

There are multiple open research issues leading to advances in technologies for storing, analyzing, sharing, citing, and integrating data. Fundamental understanding is needed not only in modeling and theory (including mathematics, statistics, and computer science) but also in new architectures, novel visualizations, and the effective utilization and optimization of computing, storage, and communications resources. All of which are motivated by research at the leading edge of science, engineering, and education. Insertion of these advances into the next generation of data infrastructure needs to occur though close collaboration between data researchers, data user communities, and the providers of current data resources.

Through its investments in foundational research NSF will seek to:

- Advance heterogeneous, large scale, distributed data collection, indexing, and management; analysis, simulation, and interpretation, at scale and in real-time; workflow, compression, and transport for high throughput systems and observatories;
- Accelerate the facilitation of data-enabled science through advances in discovery, annotation, and integration of heterogeneous datasets of varying quality, semantics, and syntax.
- Enable data visualization and analysis at large and complex scales and encourage processes and techniques to support collaboration, discovery, and the development of knowledge sharing environments;
- Stimulate research in existing disciplines and emerging interdisciplinary fields leading to new ways of expressing observations and concepts through digital media and new models for sustainability and communities of practice;
- Promote data citation and credit, to understand and reward the effort that is put into creating, curating, and sharing data; and identify metrics that measure how data is used and reused.
- Stimulate the transition to production and deployment of next generation infrastructure through diverse pathways including public-private partnerships, federal agencies, national laboratories, and the planned sunset and replacement of existing infrastructure.

3. Collaboration, Partnerships, and Grand Challenges

Investments in the National Data Infrastructure must be extensible and applicable to a wide range of disciplinary fields. But history has also shown that great scientific achievements can be made when a critical mass of data, technology, resources, and ideas are brought together in a focused and intensive effort to solve "grand challenge" problems. Addressing grand challenge problems lead to wider benefits through the extension of technologies to new areas. NSF will strategically invest in high priority areas of science and engineering through a variety of mechanisms:

- Develop a comprehensive, center scale, multidisciplinary program to address grand challenge research requiring integrative approaches to theory, experiment, data, visualization, and computation;
- Support community-building activities to promote collaboration that crosses disciplinary, institutional, and geographic boundaries through shared data concepts, standards, and semantics.
- Draw upon the latest advances in collaboration research and technology to enable communities to reach consensus on policy, standards, interoperability, research priorities, and requirements sharing. In particular, policy encouraging open, sharable data services will strongly enhance research communities' abilities to address complex grand challenges.

4. Education and Workforce Development.

The deep expertise in the collection, management, and analysis of scientific data that is needed across science and engineering will require a diverse workforce having fluency not only in computing, mathematics, and statistics but also in the collection, management, and analysis of scientific data. NSF will therefore make investments targeting these areas:

- Cross-disciplinary training in the informatics skills needed to use and advance the National Data Infrastructure;
- Facilitation and encouragement of professional career tracks in computational science and data provisioning;
- General curricula that enable data literacy and use of cyberinfrastructure tools among the public and policy making communities;
- Enhancement of student and postdoctoral training in the sciences to incorporate key concepts and techniques in the use and management of scientific data including substantive knowledge of disciplinary areas.

5. Development of data policy

Along with the development of the foundations, the tools and infrastructure, and the science, engineering, and educational needs for data-enabled science, it is also necessary to develop comprehensive policy for data, software, publications, and other digital outputs resulting from NSF-supported work. Such policy must acknowledge the new challenges, including costs surrounding support of NSF funded data repositories and data management facilities and scientific data reproducibility, storage, curation, and open dissemination in all its forms, recognizing data's importance for accelerating fundamental discovery, interdisciplinary research, and innovation in society.

Science is becoming increasingly collaborative and multidisciplinary and is enhanced when knowledge can flow easily across traditional disciplinary boundaries. Artificial barriers to the access of data and publications can slow progress and stifle opportunities for innovation. By contrast, an open networked environment will allow the evolution of knowledge and arguments to be discovered and followed through time and across disciplines, as well as verified and reproduced by other researchers. NSF needs to address new norms and practices for data citation and attribution so that data producers, software and tool developers, and data curators are credited for their contributions.

The following areas need to be addressed:

- Policy around preservation, availability, and sharing of data, publications, software, and other digital products created by NSF-supported projects.
- Citation for new forms of publication and scientific expression, including data sets, so that researchers are able to ensure their work is citable, and others are able to discover and access it.
- The linking of various forms of data, publication, software, and other digital products of research.
- Quality control and peer review of the above.

- Policy relating to the on-going financial support of NSF-funded data repositories and data management facilities, and metrics for determining sunsetting or integration with other data management structures.

The way forward

The strategic plan above defines goals and strategies for achieving NSF's vision for data in science, engineering, and education in the 21st Century. In developing an NSF-wide operations plan for data through policy, program, and synergistic activities, NSF must also consider the constraints of finite resources and competition. Priorities must be balanced between leading-edge research and production systems; between disciplinary-specific and broadly-shared infrastructure; and between grass-roots innovation and top-down standards. The goals described here can be met through coordinated efforts among NSF personnel across all units to manage a complex, diverse, and complementary set of programs in support of the data vision.