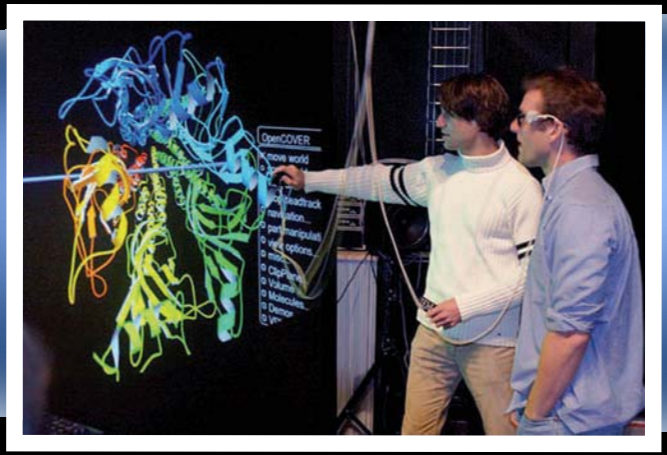
Sustainable Digital Data Preservation and Access Network Partners (DataNet)



- Background for the DataNet Program
- Context
- DataNet Partners
  - Goals
  - Characteristics
- Status

Enabling the nation's future through discovery, learning and innovation

Today's Presentation

National Science Foundation

Director

Deputy Director

National Science Board

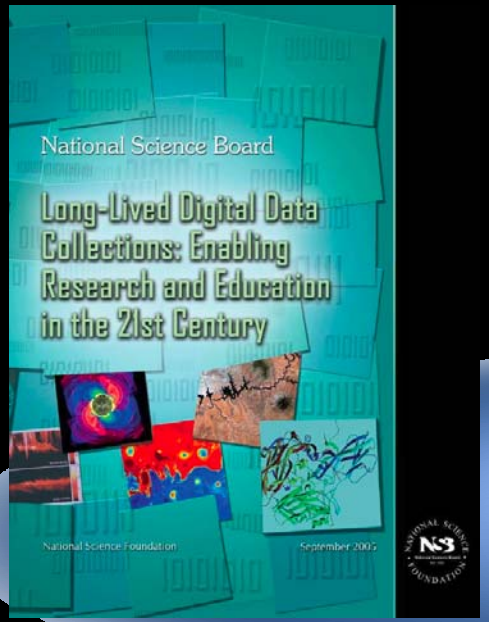
Offices

- CyberInfrastructure
- Integrative Activities
- Polar Programs
- International Science and Engineering

Research Directorates

- Biological Sciences
- Computer & Info. Science & Eng.
- Education & Human Resources
- Engineering
- Geosciences
- Mathematical & Physical Sciences
- Social, Behaviorial & Econ. Sciences

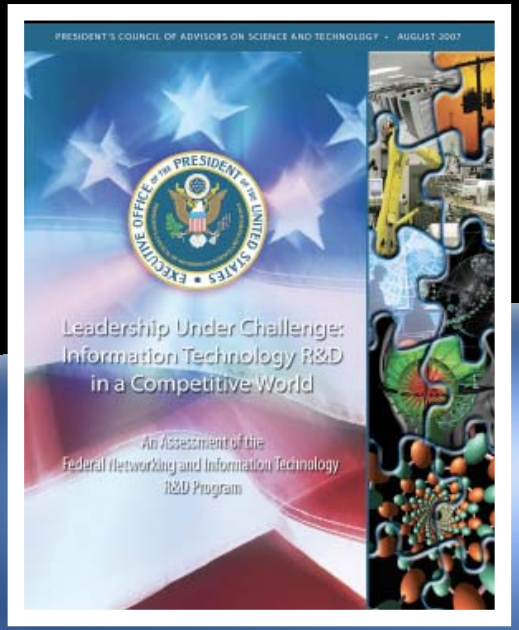
NSB Report: Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century



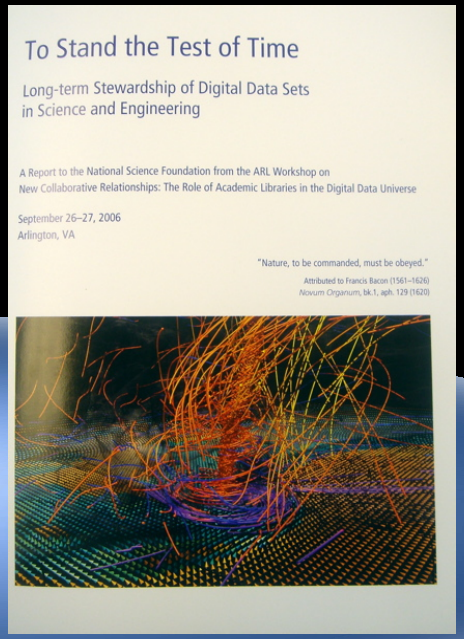
It is exceedingly rare that fundamentally new approaches to research and education arise. Information technology has ushered in such a fundamental change. **Digital data collections are at the heart of this change.** They enable analysis at unprecedented levels of accuracy and sophistication and provide novel insights through innovative information integration. Through their very size and complexity, such digital collections provide new phenomena for study. At the same time, such collections are a powerful force for inclusion, removing barriers to participation at all ages and levels of education.

National Science Board Report

PCAST Recommendations on Digital Data



NSF Supported Experts Study



Examples of Input Informing Data Activities

Storage Networking Industry Association (SNIA) *100 Year Archive Requirements Survey Report*

"The results confirmed our view that there is a pending crisis in archiving…If we're going to be able to avoid this crisis we have to create long-term methods for not only preserving information, but also for making it available for analysis in the future."

Paperwork Reduction Act*: The purposes of this subchapter are to ….

(2) Ensure the greatest possible public benefit from and maximize the utility of information created, collected, maintained, used, shared, and disseminated by or for the federal government; ….

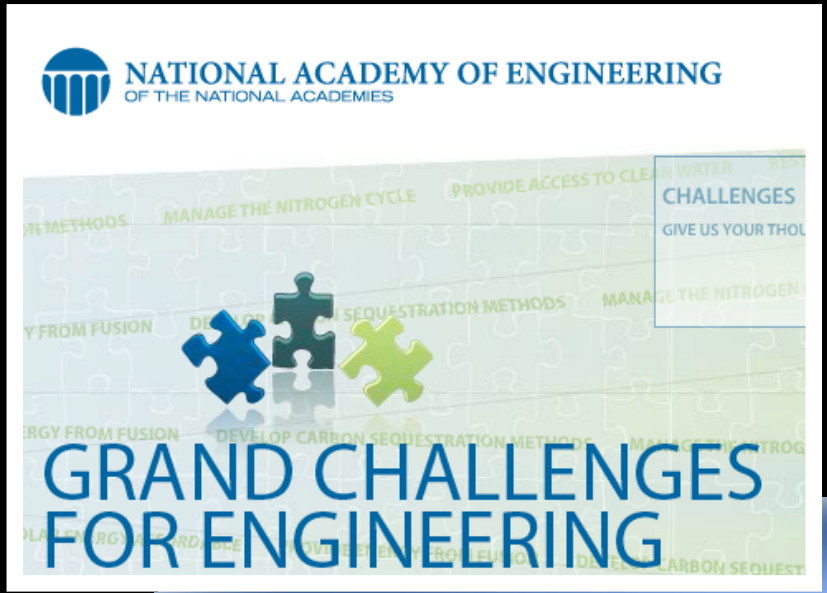
(7) Provide for the dissemination of public information on a timely basis, on equitable terms, and in a manner that promotes the utility of the information to the public and makes effective use of information technology

Office of Management and Budget (OMB) Circular A-130: Management of Federal Information Resources

Part 7. Basic Considerations and Assumptions:….

k. The open and efficient exchange of scientific and technical government information, subject to applicable national security controls and the proprietary rights of others, fosters excellence in scientific research and effective use of Federal research and development funds.

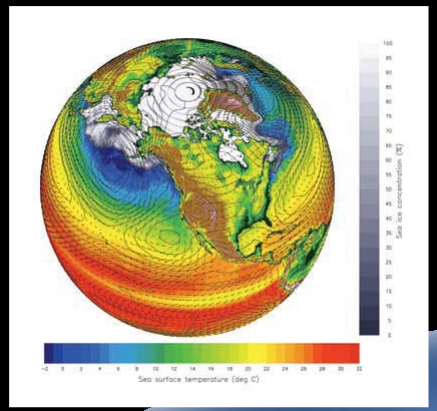
More Examples of Input Informing DataNet



http://www.engineeringchallenges.org/

Grand Challenges

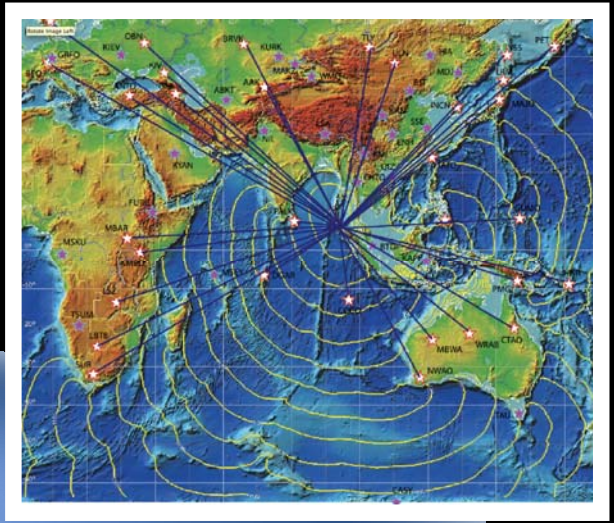
http://www.nationalacademies.org/morenews/20080312.html



- **10 Questions Shaping 21st-century Earth Science**
- How did earth and other planets form?
- What happened during earth's "dark Age" (the first 500 million years)?
- How did life begin?
- How does earth's interior work, and how does it affect the surface?
- Why does earth have plate tectonics and continents?
- How are earth processes controlled by material properties?
- What causes climate to change – and how much can it change?
- How has life shaped earth? And how has earth shaped life?
- Can earthquakes, volcanic eruptions and their consequences be predicted?
- How do fluid flow and transport affect the human environment?

Grand Challenges

- Irreplaceable Data
- Replication of Results
- Longitudindal Science and the Impact of Human Activity
- Training & Testing Models
- Enabling Interdisciplinary Science & Engineering for Grand Challenges
- Broadening Participation



Why Preserve & Share Data?

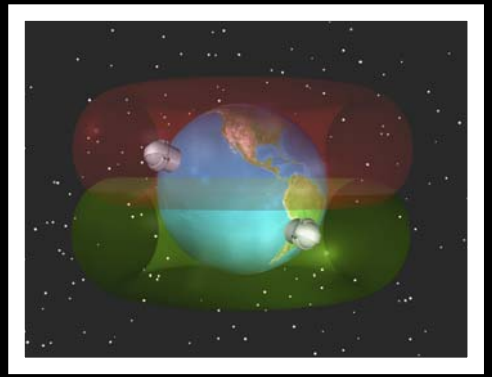
- **Irreplaceable Data**

"In 1964, the first electronic mail message was sent from either MIT, the Carnegie Institute, or Cambridge University. The message does not survive, however, and so there is no documentary record to determine which group sent the pathbreaking message."

Report of the Task Force on Archiving of Digital Information, Commission on Preservation and Access and the Research Libraries Group

Why Preserve & Share Data?

- **Irreplaceable Data**
  - **The Saga of the Lost Space Tapes**



"Maybe somebody didn't have the wisdom to realize the original tapes might be valuable sometime in the future," she said. "Certainly, we can look back now and wonder why we didn't have better foresight about this." – Dolly Perkins, Deputy Director, Goddard Space Flight Center

NASA Is Stumped in Search for Videos of 1969 Moonwalk, Marc Kaufman, Washington Post, January 31, 2007

- **WASHINGTON – NASA said today it was launching an official search for more than 13,000 original tapes of the historic Apollo moon missions.**

Seth Borenstein, Associated Press, August 15, 2006, 5:13 PM

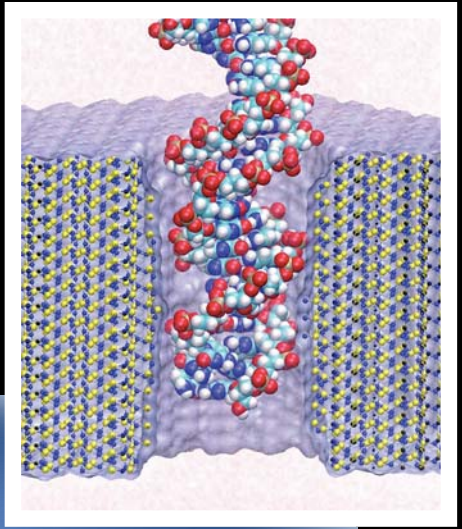
Why Preserve & Share Data?

- Irreplaceable Data

Replication of Results

"...the results of one scientist's experiment are not considered reliable until another scientist has replicated them. The reproducibility of results plays several different, crucial roles in science...[but] in many circumstances, considerations of time and money often make reproducibility impractical."

The Key Role of Replication in Science, Nancy S. Hall,
The Chronicle of Higher Education, November 10, 2000

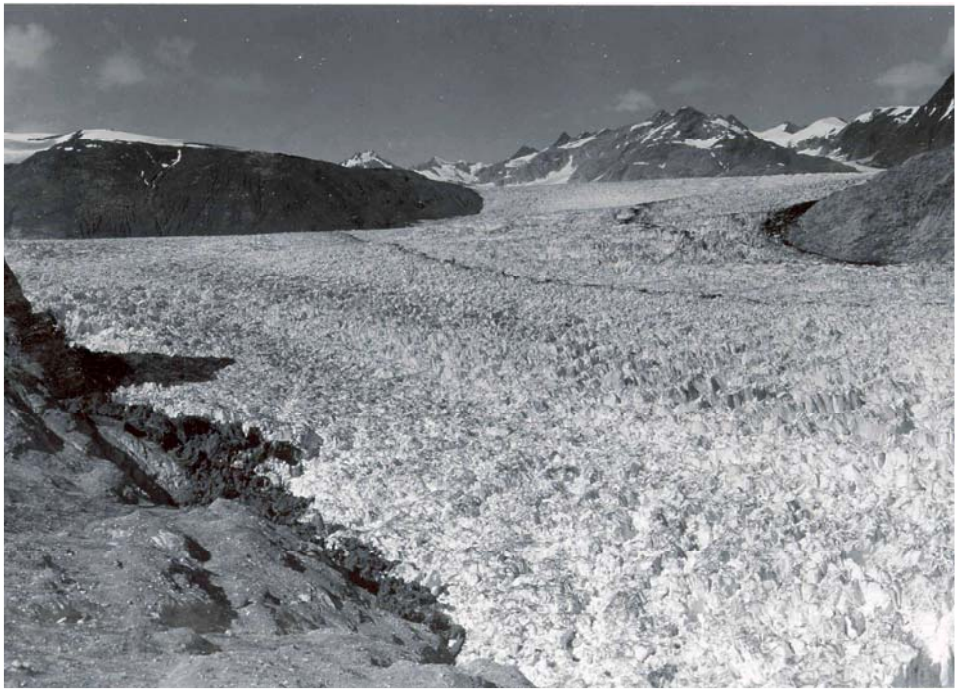


Why Preserve & Share Data?

- Irreplaceable Data
- Replication of Results
- **Data for Longitudinal Science**

With 43 years of hard-won water quality data, PI Stanley Grant observed, “When we plotted the data, our first reaction was ‘Wow! We can see the impact of the Clean Water Act!’ ”

Why Preserve & Share Data?



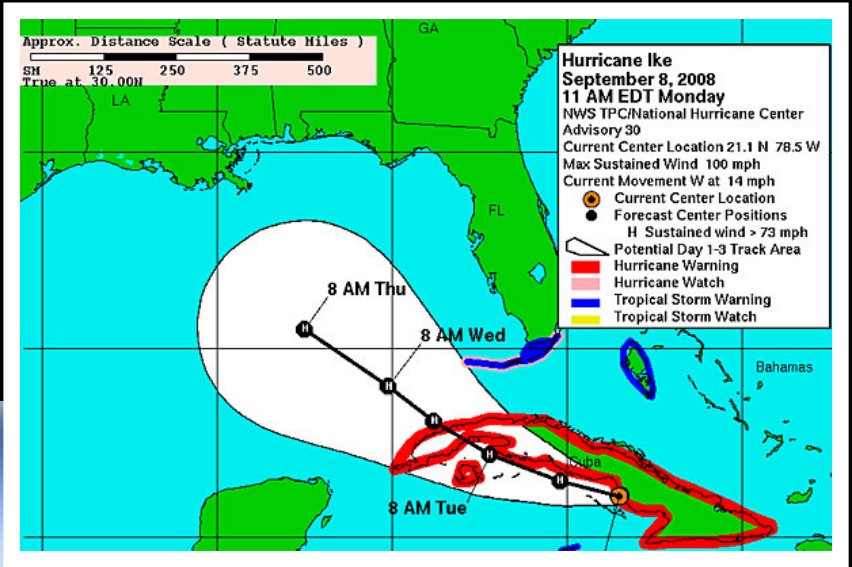
August 1941

Courtesy of Mark Parsons of the National Snow & Ice Data Center (NSIDC)

A glacier melts and becomes a lake
August 2004

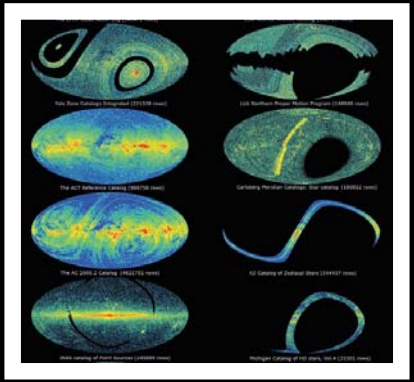
Why Preserve & Share Data?

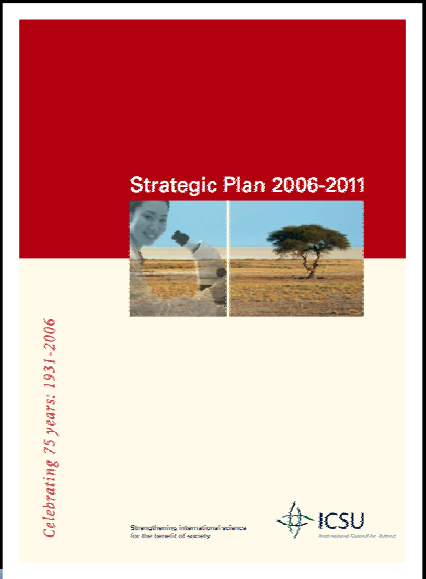
- Irreplaceable Data
- Replication of Results
- Data for Longitudinal Science
- Training and Testing Models



Why Preserve & Share Data?

- Irreplaceable Data
- Replication of Results
- Data for Longitudinal Science
- Training and Testing Models





- **Interdisciplinary Science**

"The needs and opportunities for international interdisciplinary science are greater at the turn of the 21st century than at any previous period in history... Global research problems are invariably complex and require the collaboration of many disciplines as well as many countries."

ICSU
International Council for Science

Why Preserve & Share Data?

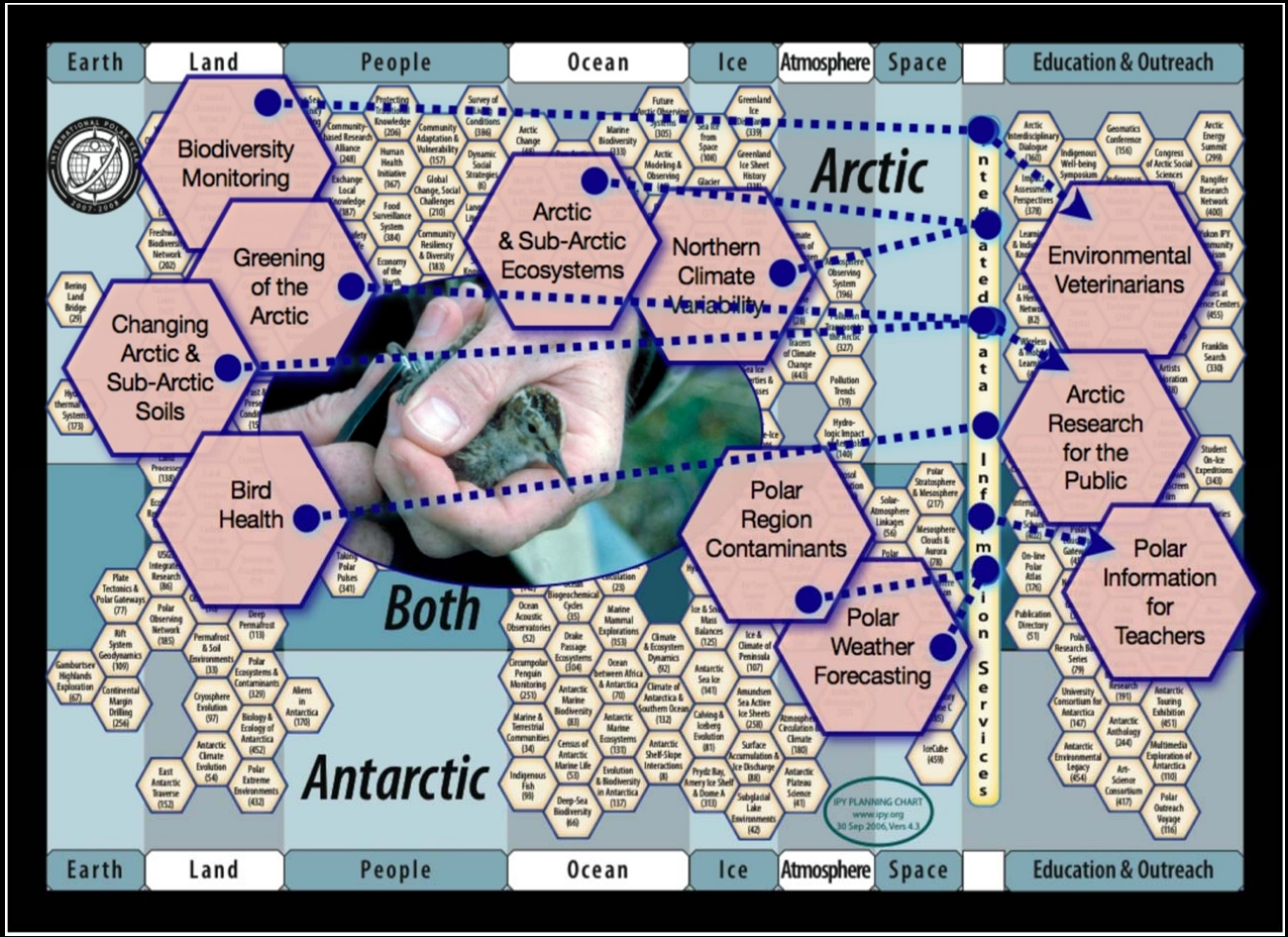
Irreplaceable Data

Replication of Results

Longitudinal Science

Training and Testing Models

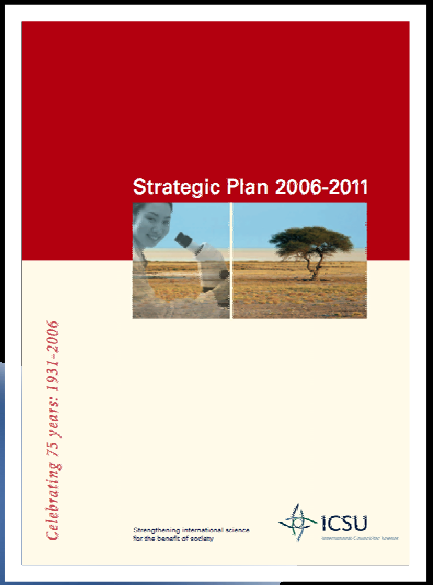
Interdisciplinary Science



Courtesy of Mark Parsons of the National Snow & Ice Data Center (NSIDC)

- Irreplaceable Data
- Replication of Results
- Data for Longitudinal Science
- Training and Testing Models
- Interdisciplinary Science

ICSU
International Council for Science

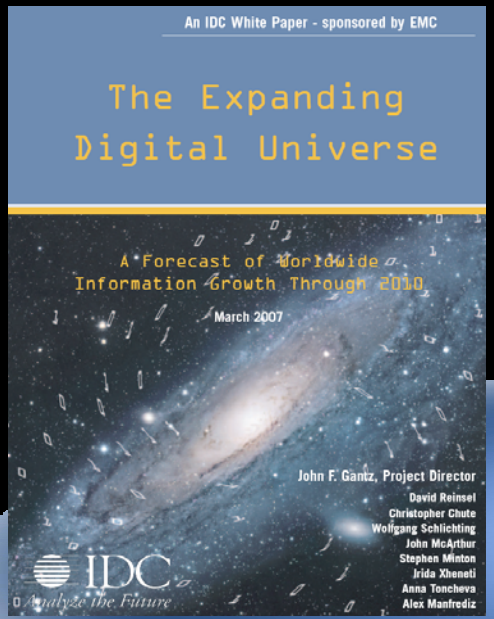


- **Broadening Participation**

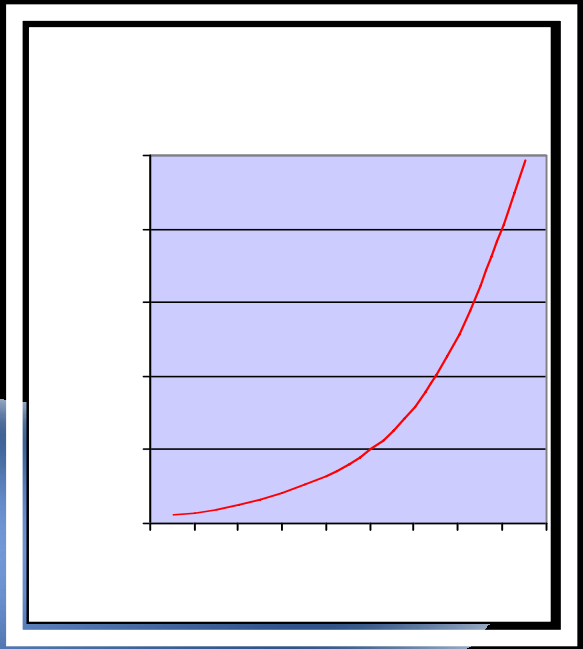
"The flow of scientific data and information is one of the most critical factors in promoting the participation of scientists...and ensuring the universality of science. As well as being of importance to science itself, publicly available scientific data are increasingly important for decision-making by governments and many sectors of society, from clinical practitioners to farmers."

Why Preserve & Share Data?

IDC - Global Marketing Intelligence Company



- "In 2007, the amount of information created will surpass, for the first time, the storage capacity available."



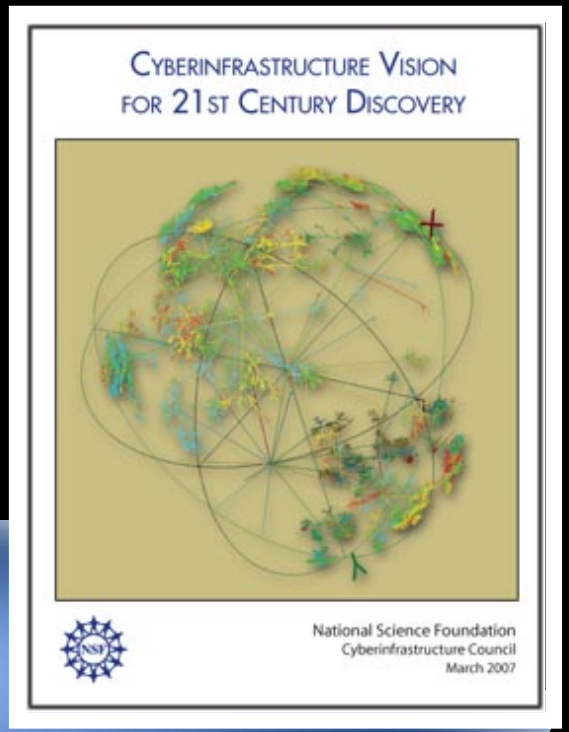
We Must Choose What to Keep

neon NATIONAL ECOLOGICAL OBSERVATORY NETWORK	~150 TB/year
LSST	~30 TB/Night
Large Hadron Collider	~15 PB/Year
NASA E·O·S·D·I·S Earth Observing System Data & Information System	~64 TB/Year
earth scope www.earthscope.org	~40 TB/Year

How Much Data?

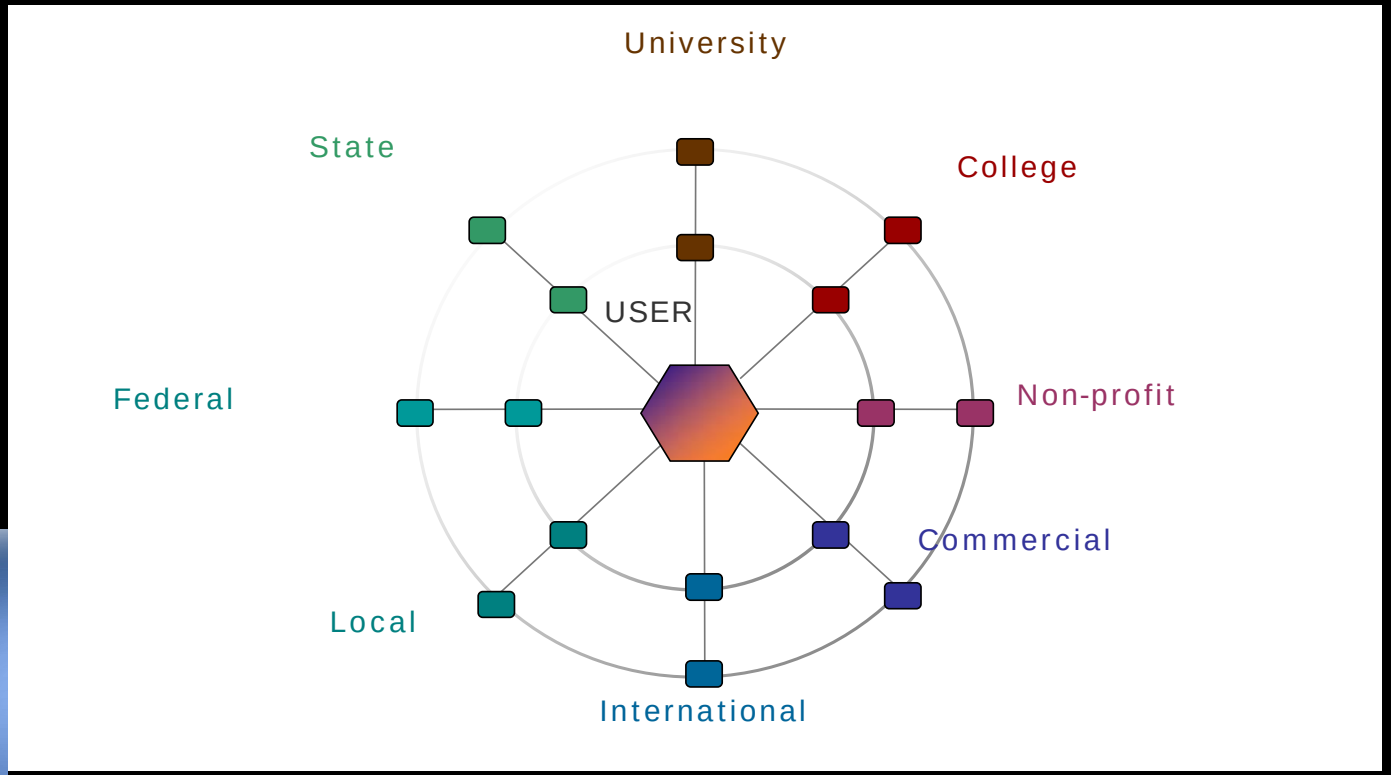
- “Science and engineering digital data are routinely deposited in a well-documented form, are regularly and easily consulted and analyzed by specialists and non-specialists alike, are openly accessible while suitably protected and are reliably preserved.”

- “Scientific visualization, including not just static images but also animation and interaction, leads to better analysis and enhanced understanding.”



NSF Cyberinfrastructure Vision for 21st Century Discovery

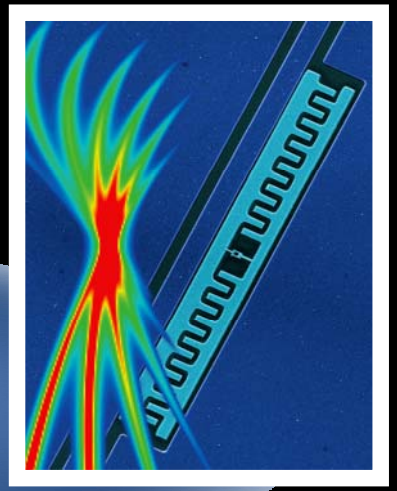
- User-centric
- Multi-Sector
- Open
- Extensible
- Evolvable
- Sustainable
- Nimble



Digital Data Preservation and Access Framework

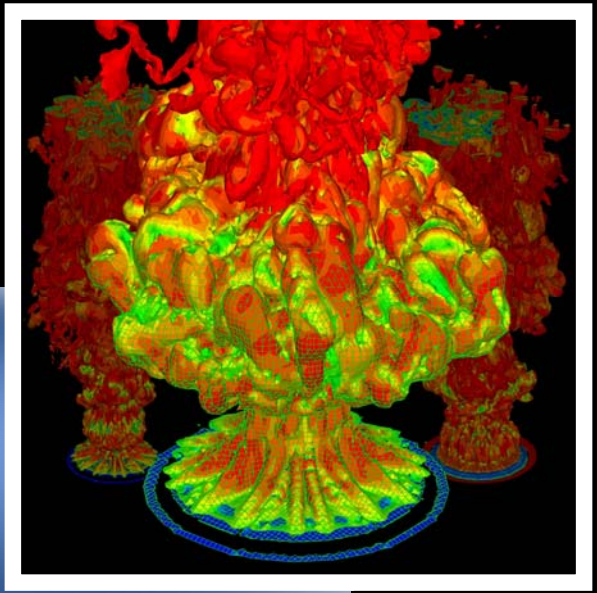
DataNet Partners: Three Primary Goals

- Achieve **long-term preservation and access** capability in an environment of rapid technology advances.
- Create systems and services that are **economically and technologically sustainable**.
- **Empower science-driven information integration** capability on the foundation of a reliable data preservation network.



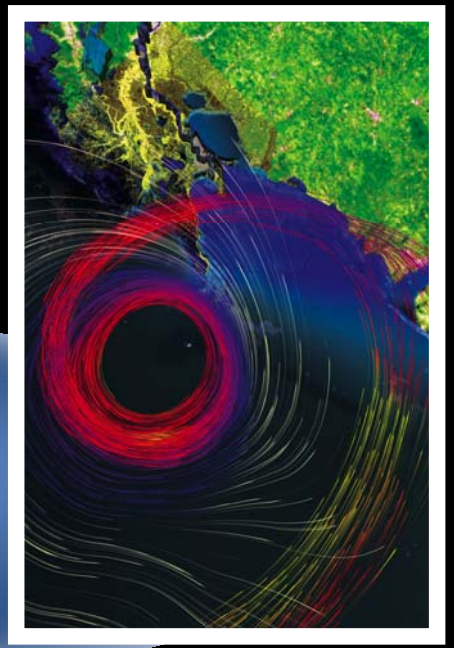
DataNet Program

- **5-6 awards planned,** 2 this year and 2-4 next year. **Each $20M ($4M/year for 5 years; potential $10M renewal.)**
- **Explore, demonstrate and understand diverse approaches** to developing in **sustainable** ways with **diverse scientific digital data content.**
- Initial focus on several disciplinary areas, with **active outreach** to more communities and more disciplines

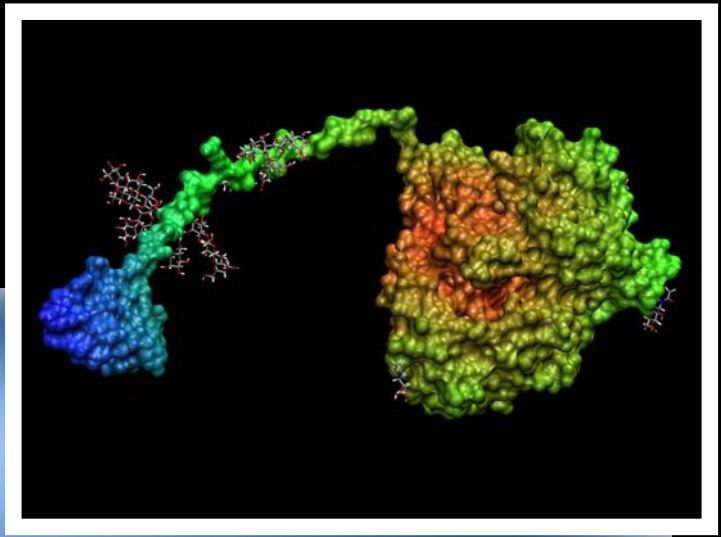


A Successful DataNet Partner Will...

- **Integrate** library and archival sciences, cyberinfrastructure, computer and information sciences, and domain science expertise.
- **Engage** at the frontiers of computer and information science and cyberinfrastructure with research and development to drive the leading edge forward.

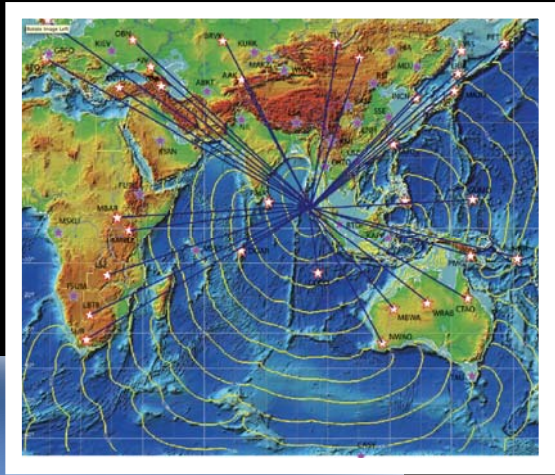


- **Provide** reliable digital preservation, access, integration, and analysis capabilities for science/engineering data over decades-long timeline.
- Continuously **anticipate and adapt** to changes in technologies and in user needs and expectations.



- Any information that can be stored in digital form and accessed electronically: numeric data, text, publications, sensor streams, video, audio, algorithms, software, models & simulations, images, etc.

- Caveats
  - Focus on data central to the scientific & engineering research & education mission of NSF.
  - Conversion to digital format is not supported by this program.
  - No support for narrowly defined, discipline-specific or institutional repositories.



Digital Data

- **Provide for full data management life cycle**
  - Data deposition/acquisition/ingest
  - Data curation & metadata management
  - Data protection, including privacy
  - Data discovery, access, use, & dissemination
  - Data interoperability, standard, & integration
  - Data evaluation, analysis, & visualization
  - Engage in research central to DataNet responsibilities
  - Education & training

- Community & user input assessment
- International engagement – collaborate & coordinate closely with preservation & access organizations to catalyze formation of a global data network
  - Foreign collaborators are expected to secure support from their own national sources.

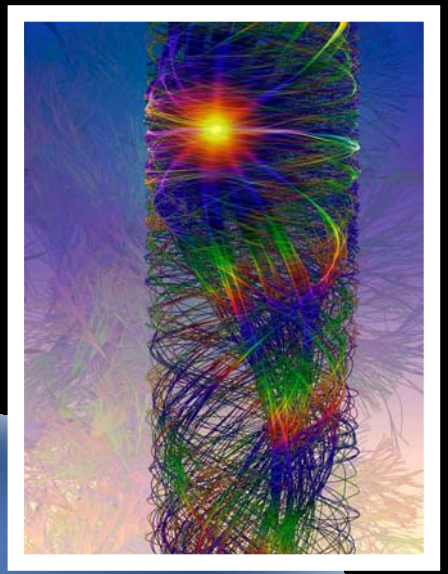
DataNet Partner Responsibilities

- **Science must drive the proposal.**
- **Start with the science drivers: What data are needed? Where do they reside now? What is needed to integrate and analyze them that is missing now?**
- **Partnerships should be based on what needs to be accomplished to support the science.**
- **Every proposal must address every DataNet requirement.**
- **Every proposal must be of interest to at least two NSF Directorates.**

Where to Begin?

DataNet Status

- Two award recommendations approved by the NSB in December: Leads are UNM and JHU
- FY09 DataNet Competition:
  - Preliminary Proposals due November 13, 2008; Panel held. Invited 7 of 23 submitted.
  - Invited Full Proposals due May 15, 2009
  - Site Visits planned for August 2009



Led by JHU Libraries, with Associate Dean for Libraries as PI.
Building on JHU Library success with Sloan Digital Sky Survey and National Virtual Observatory, plus success of partners.
Besides astronomy, initial focus on observational data about turbulence, biodiversity, and environmental science.
Especially suited to terabyte-scale data sets, but with strong focus on data from "the long tail of small science."

Data Conservancy

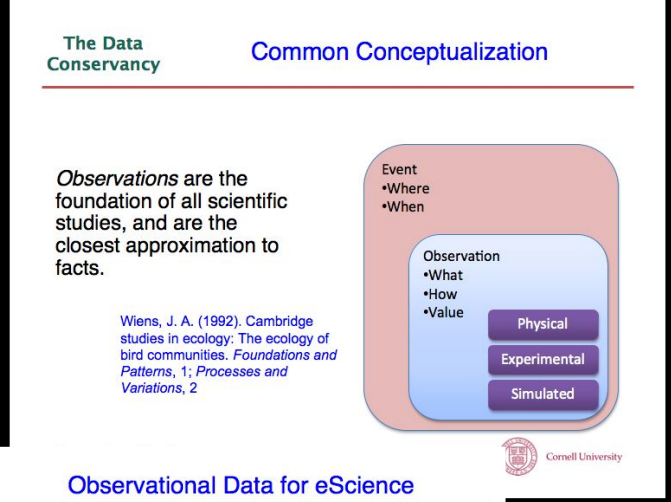
Redefine the role of university libraries to include a leading role in digital preservation.

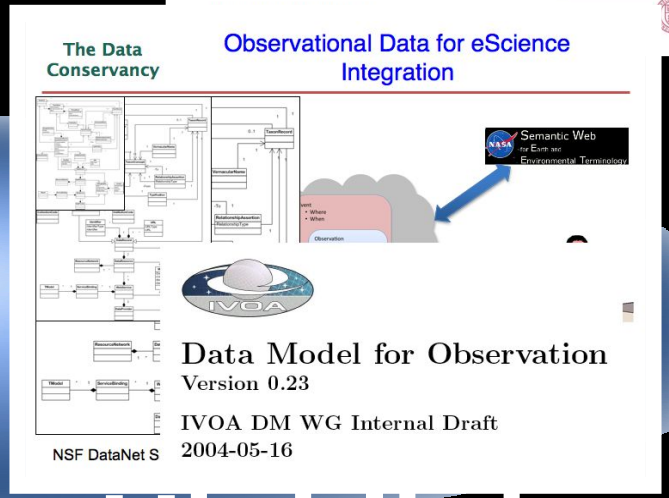
User/task-centric approach to requirements development and system design.

Develop a unified model for observational data that works across disciplines, with data analysis tools and methods.

Change the culture and practice of scholarly communication by working with publishers, incentivizing data deposition & re-use.

Extend the curriculum in Library and Information Science to address data curation and preservation.





Data Conservancy Highlights

Focus on observational data and enable scientists to identify causal and critical relationships between physical, biological, ecological, and social systems. Sample scientific problems:

Relationships among intense rainfall, seismicity, land-use practices, and landslides, especially how the interactions increase the risk of landslides, endangering lives and property.

Global, national and regional drivers of land and energy use in mega-cities and the corresponding impacts on the carbon cycle and climate change.

Drivers for the global food crisis, including impacts of bio-fuel production and various socio-political factors within countries.

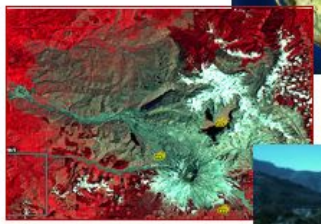
Data Conservancy

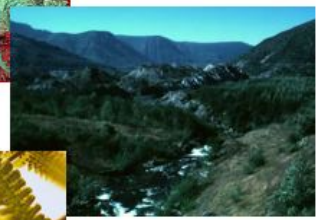
Trailblazing effort to establish a library-based scientific digital data preservation cyberinfrastructure paradigm.
User-centered methodology conducting research to understand science and engineering data practices, with focus on informing system requirements.
Comprehensive plan for curriculum development in data management and data science for schools of library and information science.

Data Conservancy

DataNetONE Vision

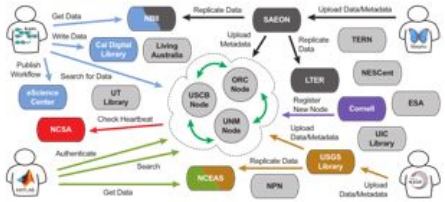
- To enable the long-term access and use of preserved earth observation data
  - Data are diverse and complex (multi-scale, multi-discipline, and multi-national)
- Agile, evolutionary development building upon a progressively refined, robust core framework





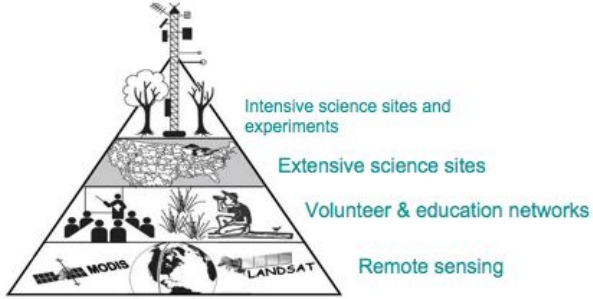
DataNetONE will ***enable new science and knowledge creation through universal access to data about life on earth and the environment that sustains it by:***

- making the scientist an active member of the data preservation process,
- creating cyberinfrastructure that supports the full data life cycle,



- promulgating cultural changes that value data stewardship and data sharing,
- broadly promoting best practices, and

- engaging citizens in science.



DataNetONE

(Observation Network for Earth)

Integrating earth observing networks, both existing and under development.
Strong emphasis on user community engagement via working groups.
Focus on creating cultural change in the user community to promote data deposition and re-use.
Strong disciplinary roots, with project led by a biologist who is an ecological researcher at UNM

DataNetONE

Driven by research challenges in climate change and biodiversity, with aim to improve ability to observe, model, assess and adapt to impacts of climate change, particularly on a regional scale, and assure availability of critical long-term climate data.

Sample scientific problems:

How will crop pest infestation patterns change in the next 10 years and what willbe the impact on maximum food yield?

How will climate change throughout the Midwest impact times of flowering and harvest?

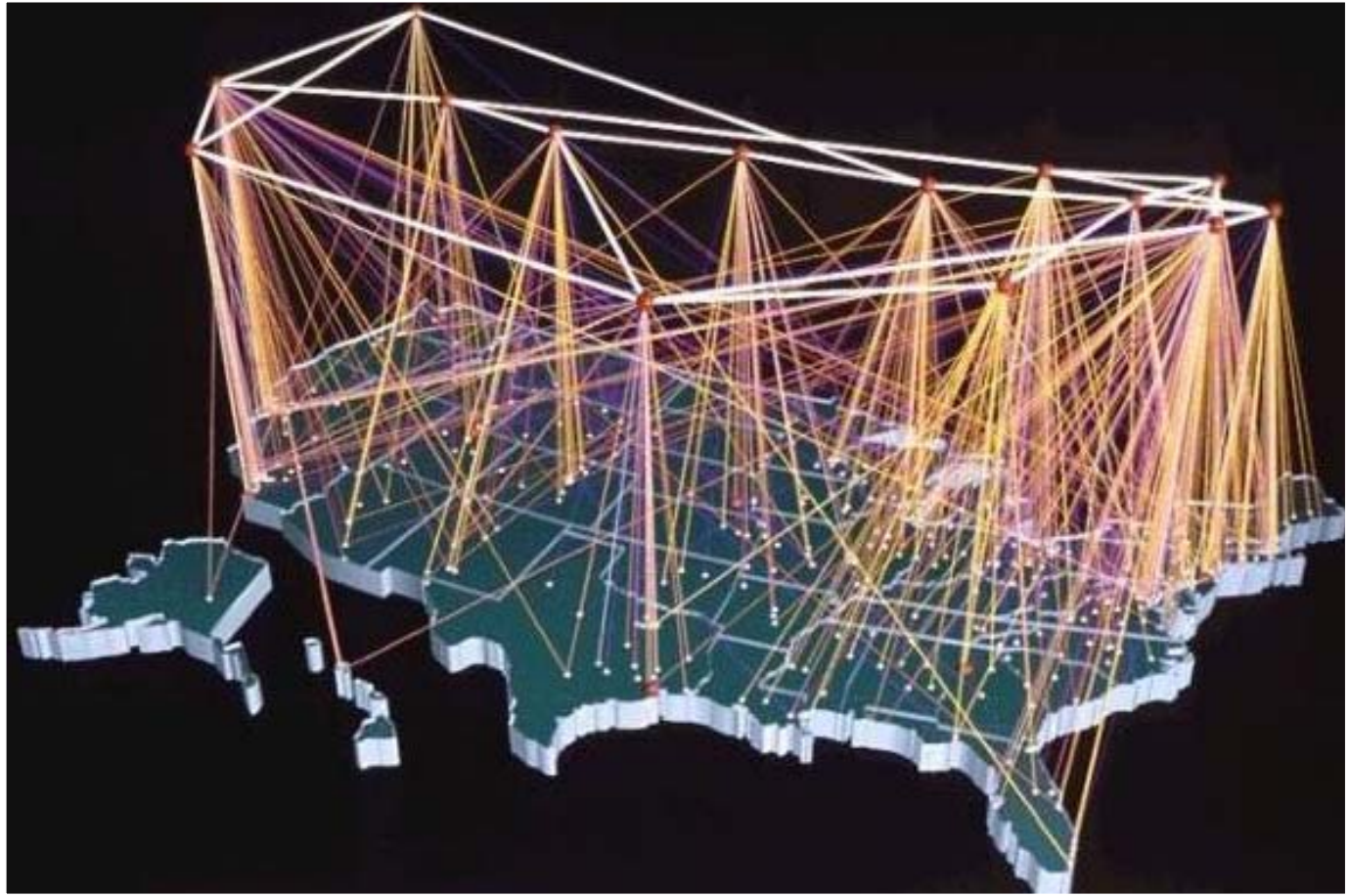
What are the relationships among human population density, atmospheric nitrogen and carbon dioxide, energy consumption, and global temperatures?

DataNetONE

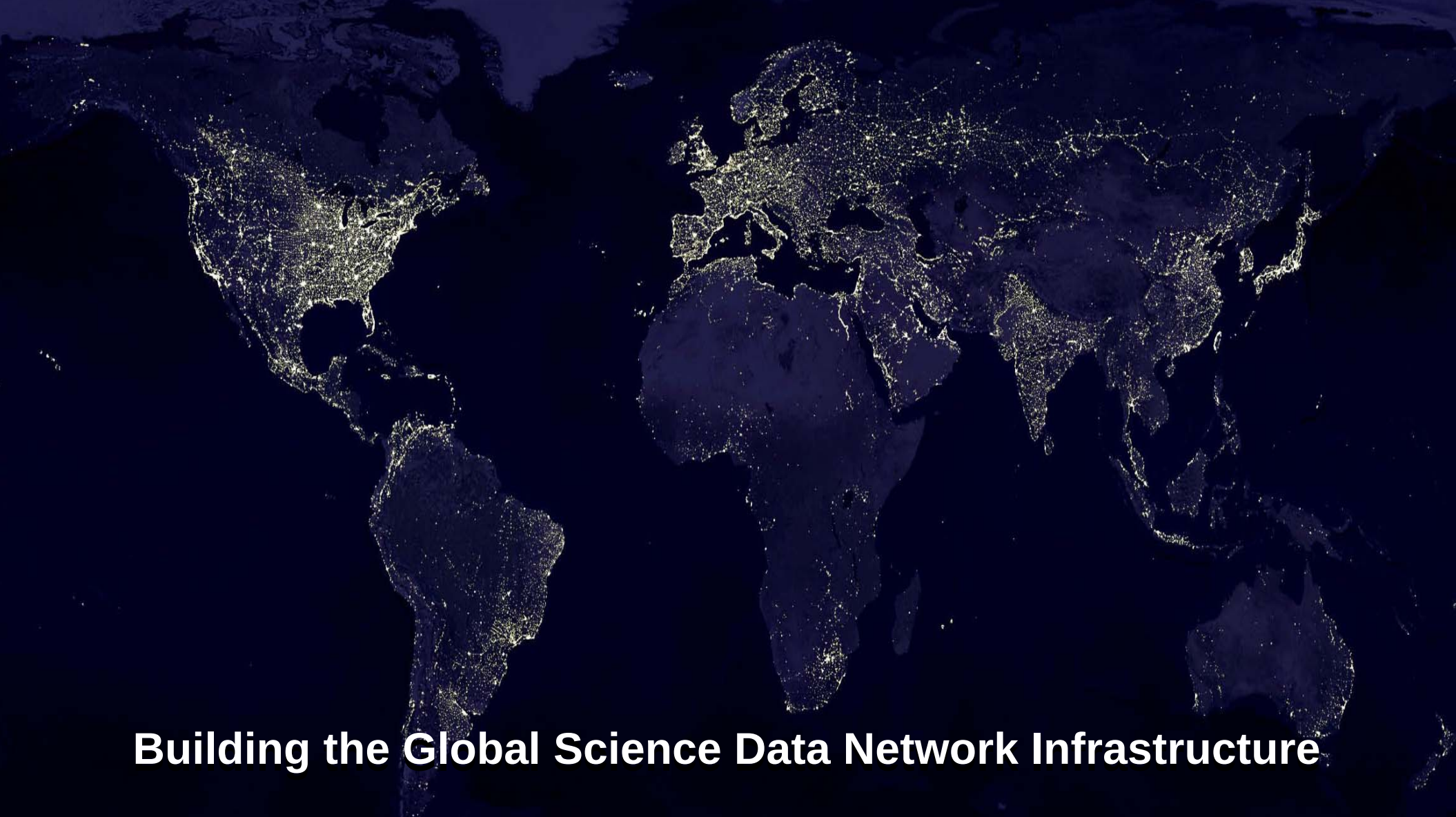
(Observation Network for Earth)

Each focuses on observational data, using complementary approaches to research and organizational structure.
Each can capitalize on results of the other.
Differences in approach provide an element of risk management.
Sufficient overlap in science drivers to motivate building cooperative relationships and interoperability, with significantly greater scientific impact as a result.

Synergistic Awards



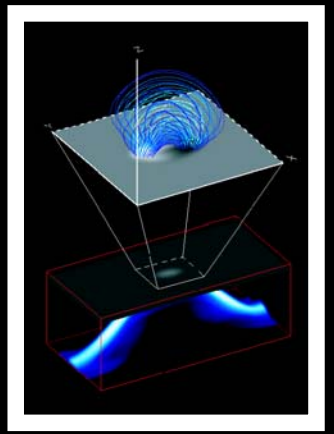
Goal: A DataNet Partners Network of Networks

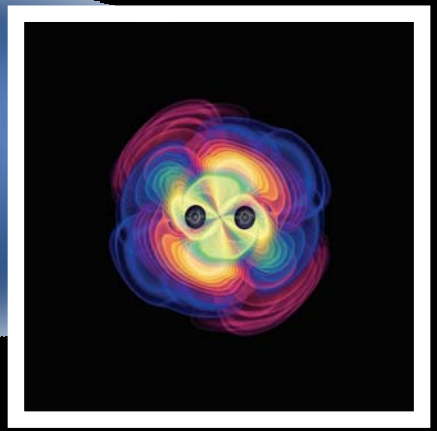


Building the Global Science Data Network Infrastructure

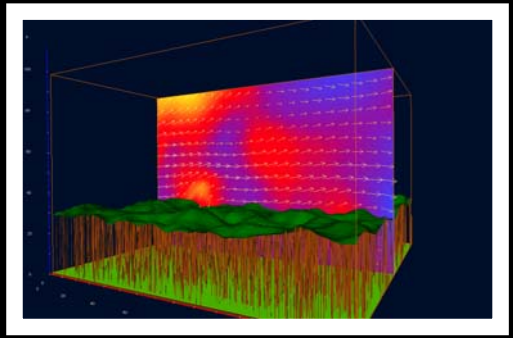
Critical Needs & Opportunities

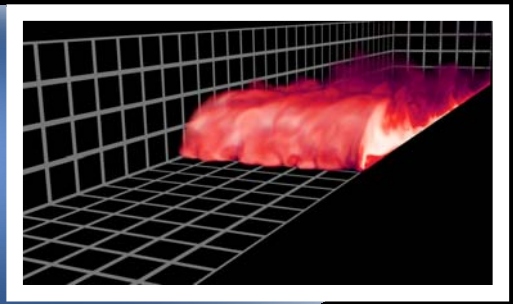
- Extending/evolving collections & services of DataNet Partners
- Outreach
  - Changing the culture, practices & reward structures
  - Promote reuse & repurposing of data
- Support for interdisciplinary grand challenge research
- Interoperability across collections



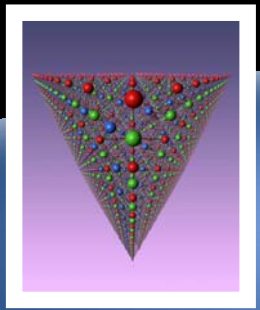


- **Integrating** DataNet Partners with other CI activities
- Hardware & software architectures for **data-intensive computing**
- **Information synthesis**, data mining & analytic tools
- **Broaden participation** via data sharing/re-use
  - Geographically
  - K-12, Undergraduate, etc.
- **Sustaining** DataNet Partners for Phase 2
- Extend DataNet Partners beyond the first 5-6 awards





- **“Without investments made by previous generations, we would not enjoy the seemingly invisible infrastructure that makes our modern lives possible. It goes without saying that if we don’t make similar investments now we will rob future generations of the quality of life they should enjoy.”**



Princeton University Engineering School, Feb. 19, 2008, Greatest Technological Research Challenges of the 21st Century Identified by Expert Panel. *Science Daily*

http://www.sciencedaily.com/release/2008/02/080215151157.htm

Greatest Technological Challenges of the 21st Century

Thank you!