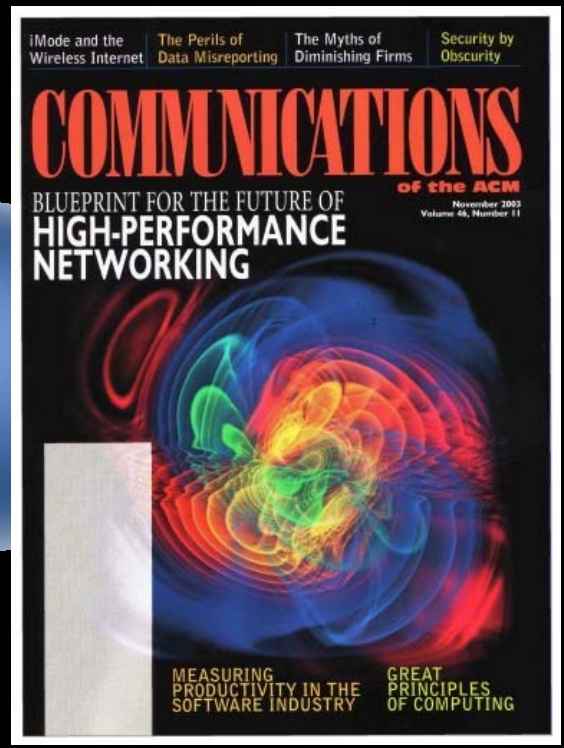
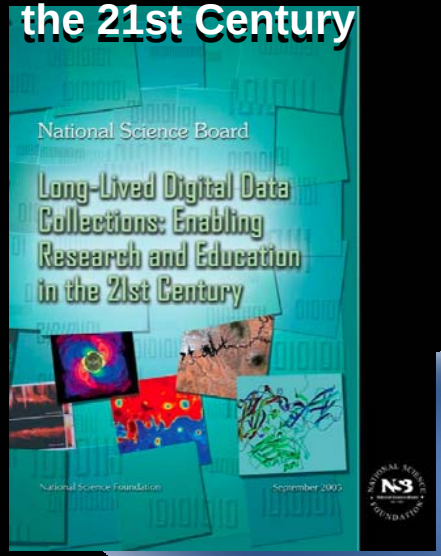
NSF's Data-Driven Vision for Cyberinfrastructure

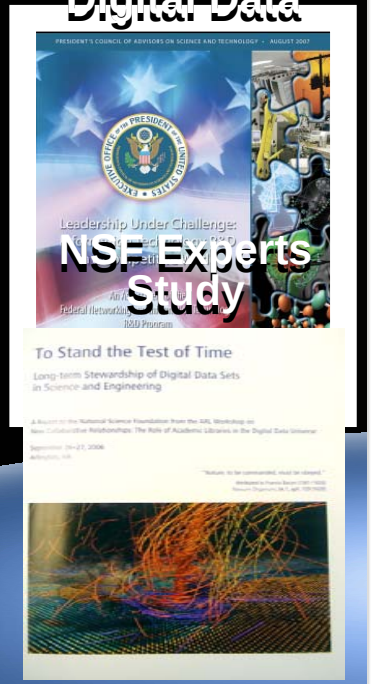
Edward Seidel
Director, Office of Cyberinfrastructure
National Science Foundation
hseidel@nsf.gov



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



PCAST Digital Data



Industry

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

"there is a pending crisis in archiving… we have to create long-term methods for preserving information, for making it available for analysis in the future." 80% respondents: > 50 yrs; 68% > 100 yrs

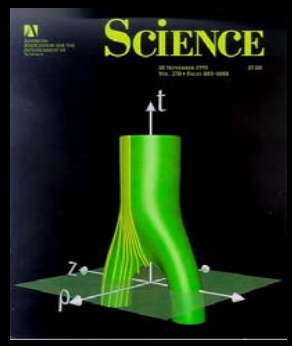
Wired, Nature

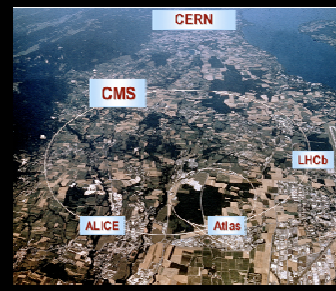
"The data deluge represents an opportunity to advance U.S. leadership in science and technology, and harnessing it has become a national priority"

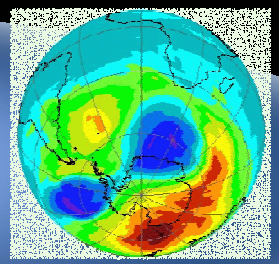
Edward Seidel
hseidel@nsf.gov

Data-Driven Collaborations for Complex Problems
A Great Challenge of 21st Century

- Every field of science
  - General Relativity
  - High Energy Physics
  - Geosciences, Bio, SBE...
  - And all combinations...
- Science and Society being transformed by CI- and Data-Driven approaches
- CI plays central role in data-driven collaborations for complex problems
  - No single community can attack challenges
  - Technical and social issues
  - Data intensive (“4th paradigm”)

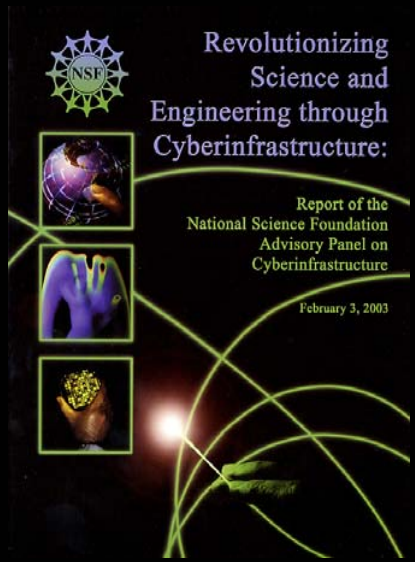




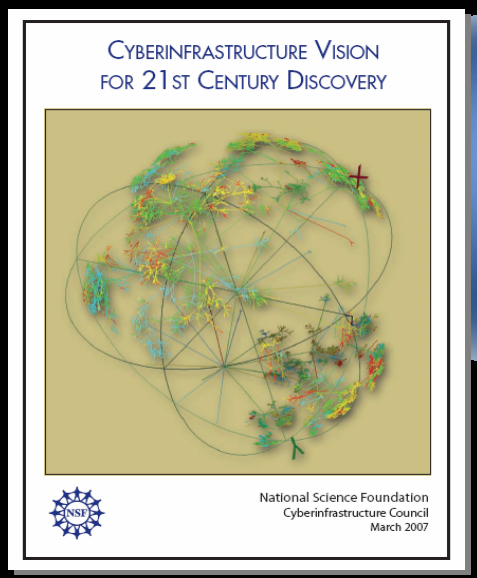


Edward Seidel
hseidel@nsf.gov
Office of Cyberinfrastructure

NSF Vision



"national-level, integrated system of hardware, software, data resources & services... to enable new paradigms of science"



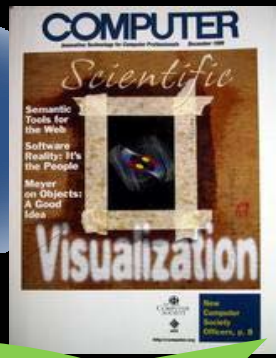
1. Virtual Organizations for Distributed
2. High Performance Computing
3. Data & Visualization/Interaction
Opportunities

Edward
hseidel@nsf.
ffice of
berinfrastructure

Data, Data Analysis, and Visualization

- "Any cogent plan must a
  phenomenal growth of d s
- Goals are to
  - Catalyze the development of a system of science and engineering data collections that is open, extensible, and evolvable
  - Support development of a new generation of tools and services for data discovery, integration, visualization, analysis and preservation
- The resulting national digital data framework will be an integral component in nati

Where is my simulation data? Compare with experiment? Publication?



I can't visualize my results!

Edward Seidel
hseidel@nsf.gov

Examples of Science Drivers for OCI Data Activities

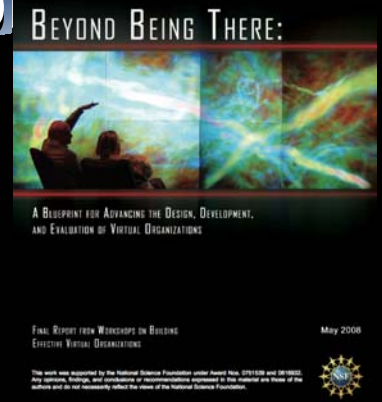
Edward Seidel
hseidel@nsf.gov

Issues of Major Concern

- How to build data-driven science?
  - What are the methodologies? How do we effect culture shift? How do we develop data-scientists? How do we create cross-disciplinary data-communities?
- Technological and economic issues
  - How do we store, mine, curate digital information?
  - How do we sustain (and pay for) the activities?
- Open Access
  - Data collections, software, publications...
  - We need to change the paradigm

Ideas for Rough Plan for OCI

- Attempt to fill out holes in current CI coverage
  - *Data*, software, algorithms, networks...
  - Alternate computing modalities beyond HPC: Clouds, data, campus integration, etc...
  - Task Forces: Data, 5 others...
- Workforce development!
- OCI as base for CI development, deployment & *computational science* research and education
- New programs: VOs attack complex pro
  at frontiers of science and engineering
  - CI-driven, with above pieces in place first
  - This is the heart of computational science!



Edward Seidel
hseidel@nsf.gov