Revolutionizing Science & Engineering: The Role of Cyberinfrastructure

UCSD Jacobs School of Engineering Research Review
February 28, 2003

Peter A. Freeman

Assistant Director of NSF

For Computer & Information Science & Engineering (CISE)

"[Science is] a series of peaceful interludes punctuated by intellectually violent revolutions… [in which]… one conceptual world view is replaced by another."

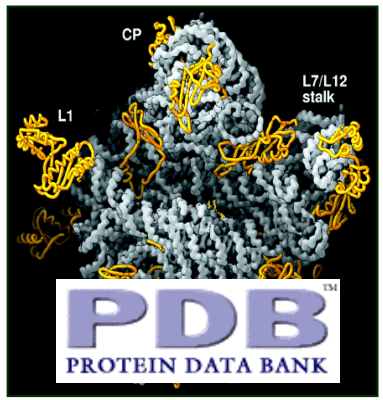
--Thomas Kuhn

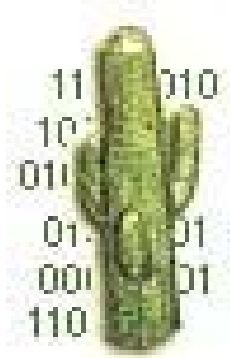
The Structure of Scientific Revolutions

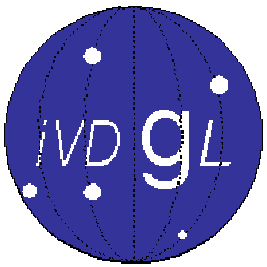
Current technology and innovative computational techniques are **revolutionizing** *science and engineering.*

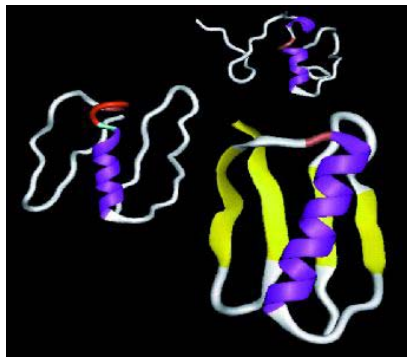
- Takeaways are:
  - A true revolution is occurring in science and engineering research
  - It is driven by computer technology
  - Data is paramount
  - There are many technical challenges for computer science and engineering

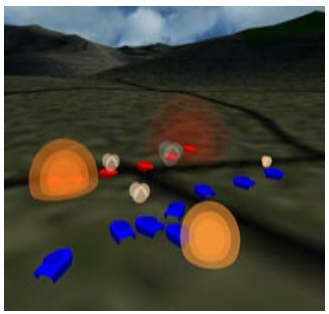
Information Infrastructure is a First-Class Tool for Science Today

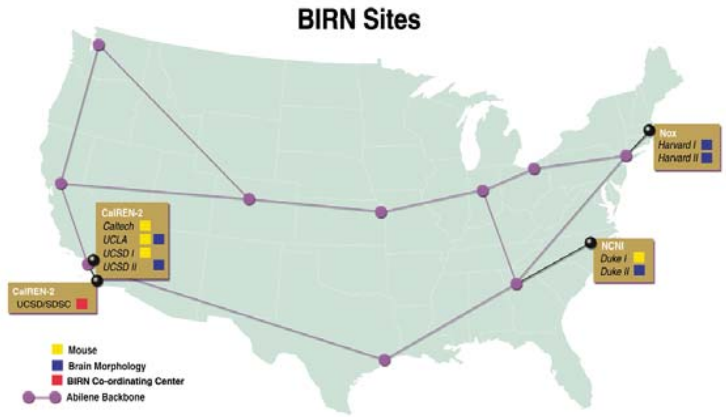












The Information Tsunami --An Example

- **Terabyte [1,000,000,000,000 bytes OR 10^{12} bytes]**
- 1 Terabyte: An automated tape robot OR all the X-ray films in a large technological hospital OR 50000 trees made into paper and printed OR daily rate of EOS data (1998)
- 2 Terabytes: An academic research library OR a cabinet full of Exabyte tapes
- 10 Terabytes: The printed collection of the US Library of Congress
- 50 Terabytes: The contents of a large Mass Storage System
- 400 Terabytes: National Climactic Data Center (NOAA) database
- **Petabyte [1,000,000,000,000,000 bytes OR 10^{15} bytes]**
- 1 Petabyte: 3 years of EOS data (2001), OR 1 sec of CMS data collection
- 2 Petabytes: All US academic research libraries
- 8 Petabytes: All information available on the Web
- 20 Petabytes: Production of hard-disk drives in 1995
- 200 Petabytes: All printed material OR production of digital magnetic tape in 1995
- **Exabyte [1,000,000,000,000,000,000 bytes OR 10^{18} bytes]**
- 2 Exabytes: Total volume of information generated worldwide annually
- 5 Exabytes: All words ever spoken by human beings
- **Zettabyte [1,000,000,000,000,000,000,000 bytes OR 10^{21} bytes]**
- **Yottabyte [1,000,000,000,000,000,000,000,000 bytes OR 10^{24} bytes]**

The Future: multi-petabyte databases generated by new science programs

Science program/application	2002 archive size by year in terabytes	2003	2006
Large Hadron Collider (LHC)	100	500	2500
National Virtual Observatory (LHC)	35	55	1000
Laser Interferometer Gravitational Wave Observatory	20	100	600
Neuroscience Imaging	<1	50	200
Network for Earthquake Engineering Simulation	<1	5	50
National Ecological Observatory Network	<1	5	50

Source: PACI

The Changing Style of Observational Astronomy

The Old Way:	**Now:**	**Future:**
Pointed, heterogeneous observations (~ MB - GB)	Large, homogeneous sky surveys (multi-TB, ~ 10^6 - 10^9 sources)	Multiple, federated sky surveys and archives (~ PB)
Small samples of objects (~ 10^1 - 10^3)	Archives of pointed observations (~ TB)	**Virtual Observatory**

The US National Virtual Observatory

•Astrophysics Decadal Survey Report recommended the creation of the NVO as the highest priority in their “small projects” category.

•Federated datasets will cover the sky in different wavebands, from gamma- and X-rays, optical, infrared, through to radio.

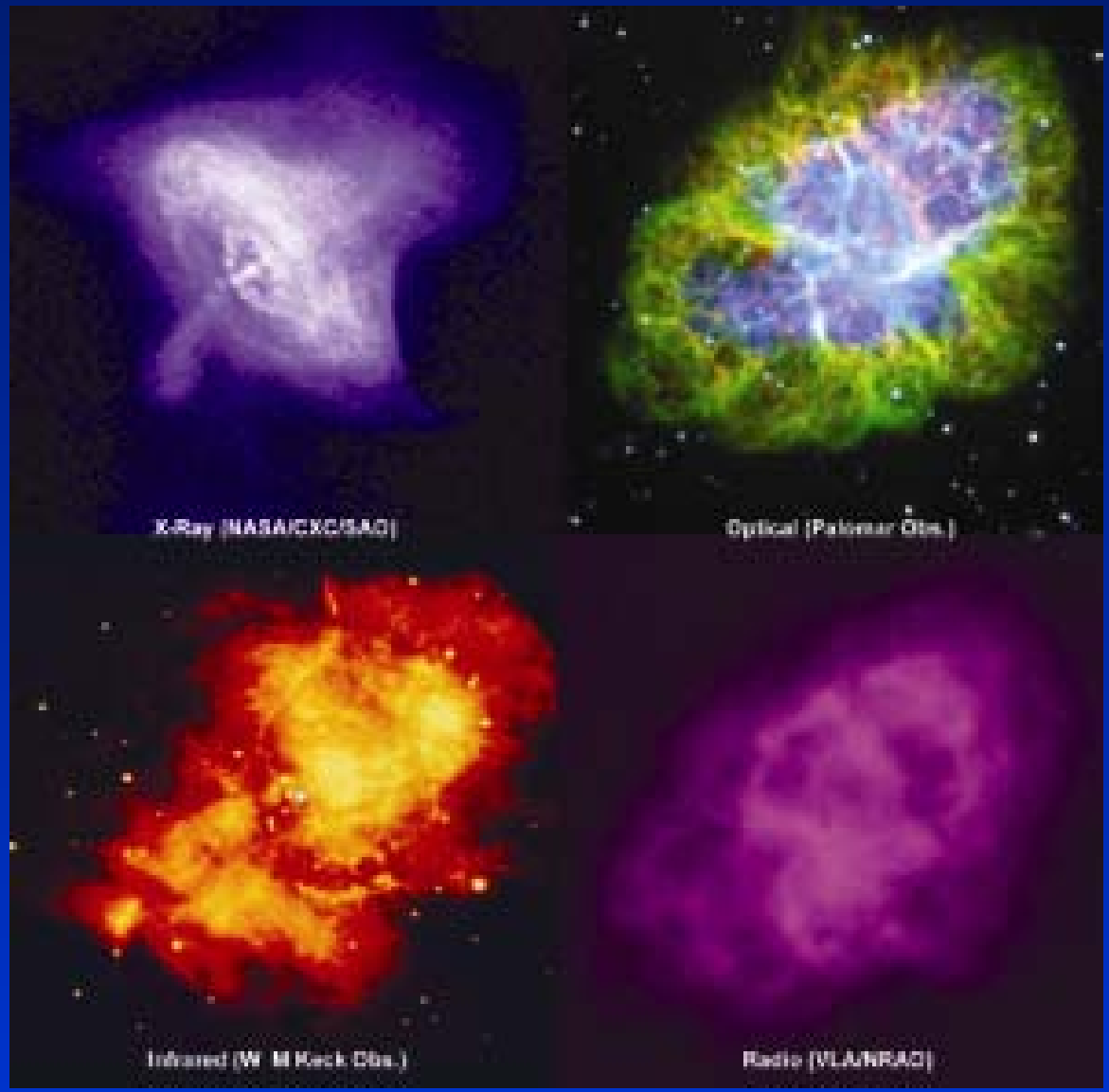
•Catalogs will be interlinked, query engines will become more and more sophisticated, and the research results from on-line data will be just as rich as that from "real" telescopes.

•Planned Large Synoptic Survey Telescope will produce over 10 petabytes per year by 2008.

•These technological developments will fundamentally change the way astronomy is done.

Crab Nebula in 4 spectral regions

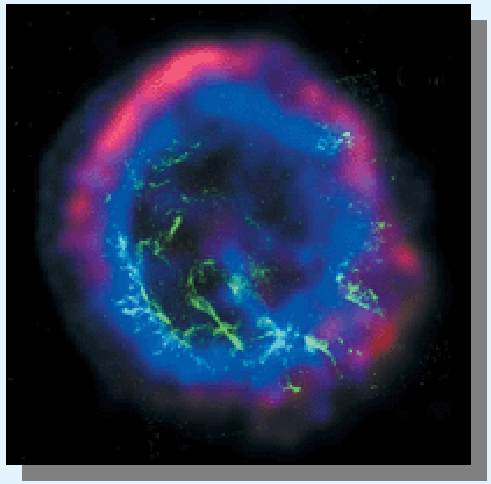
X-ray, optical, infrared, radio



Data Sharing Today – Downloading the Heavens

National Virtual Observatory

- **NVO combines over 100 TB of data from 50 ground and space-based telescopes and instruments to create a comprehensive picture of the heavens**
  - Sloan Digital Sky Survey, Hubble Space Telescope, Two Micron All Sky Survey, National Radio Astronomy Observatory, etc.
- **Astronomy community came together to set standards for services and data**
  - Interoperable, multi-terabyte online databases
  - Technology-enabled, but science driven.



Supernova remnant in the Small Magellanic Cloud, (satellite galaxy of the Milky Way)

image is composite from three data sources made possible by NVO

Hardware Systems at SDSC

- 6 Petabyte archive
- 60 Terabyte disk cache.
- 64-processor SF15k data analysis platform.
- 64-processor SF15k Storage Resource Broker data management platform

Collections Managed at SDSC

- Storage Resource Broker Data Grid*
  - 40 Terabytes
  - 6.7 million files
- BIRN Data Grid
- PDB

* including collections for:

- 2Micron All Sky Survey
- Digital Palomar Sky Survey
- Visible Embryo digital library
- HyperSpectral Long Term Ecological Reserve data grid
- Joint Center for Structural Genomics data grid
- Scripps Institution of Oceanography exploration log collection
- SIO GPS and environmental sensor data collections
- Transana education classroom video collection
- Alliance for Cell Signaling micro-array data
- NPACI researcher-specific data collections
- Hayden Planetarium data grid

Data Projects at SDSC

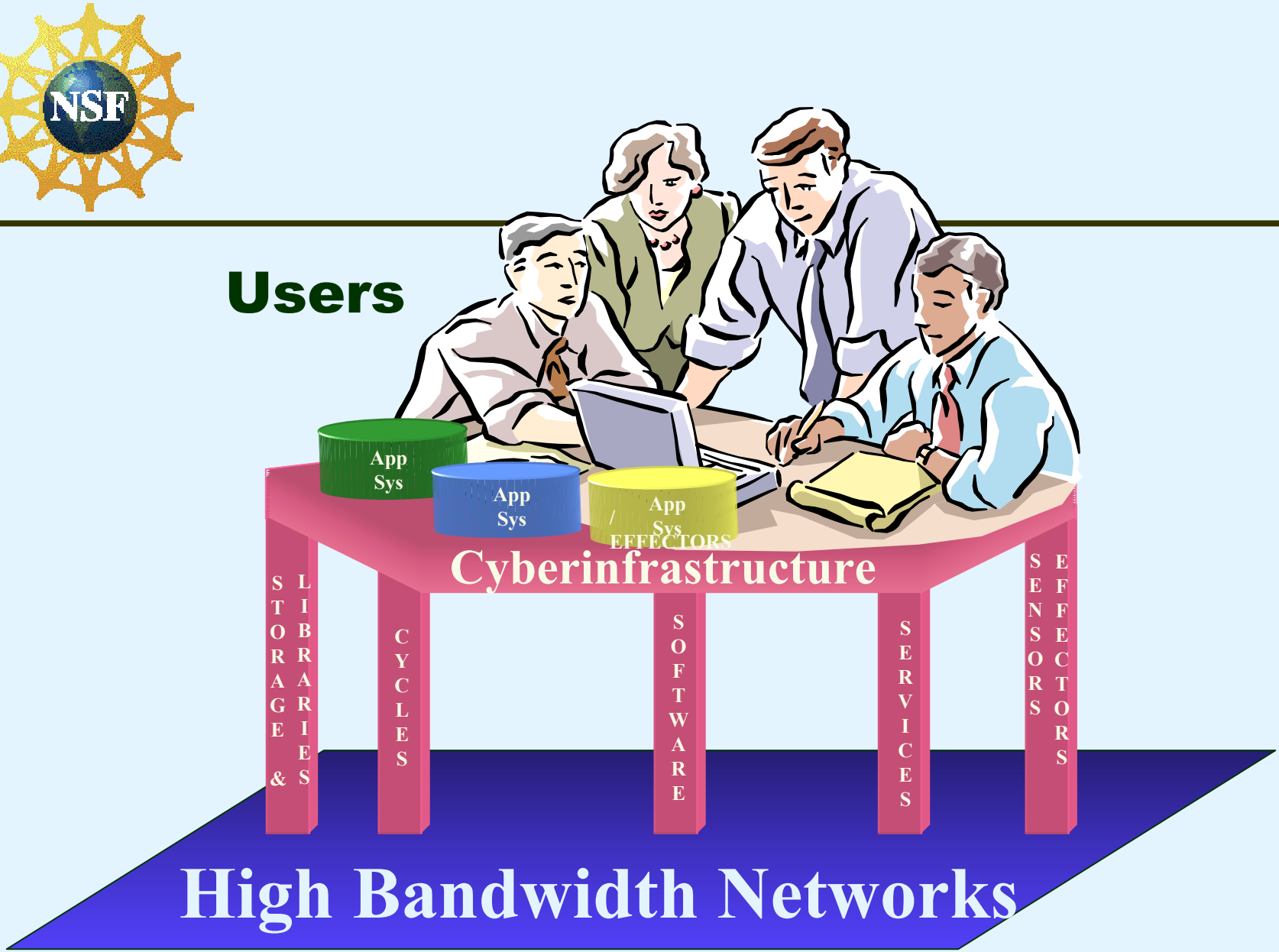
- Projects include:
  - NSF Grid Physics Network ($625,000 to UCSD, through 6/30/05)
  - NSF Digital Library Initiative ($750,000 through 2/28/04)
  - NARA supplement to NPACI ($2,100,000 through 5/31/05)
  - DOE Logic-based data federation ($546,000 through 8/14/04)
  - DOE Particle Physics Data Grid ($472,000 through 8/14/04)
  - DOE Portal Web Services ($469,000 through 5/31/05)
  - NSF National Virtual Observatory ($390,000 through 9/30/06)
  - NSF Southern California Earthquake Center ($2,714,000 to UCSD through 9/30/06)
  - NSF National Science Education Digital Library ($765,000 through 9/30/06)
  - Library of Congress ($74,500 through 4/30/03)
  - NIH BIRN

Converging Trends

- Transformative power of computational resources for S&E research
- Recognition of the importance of computation to S&E and of S&E to the Nation
- Power and capacity of the technology

Cyberinfrastructure consists of …

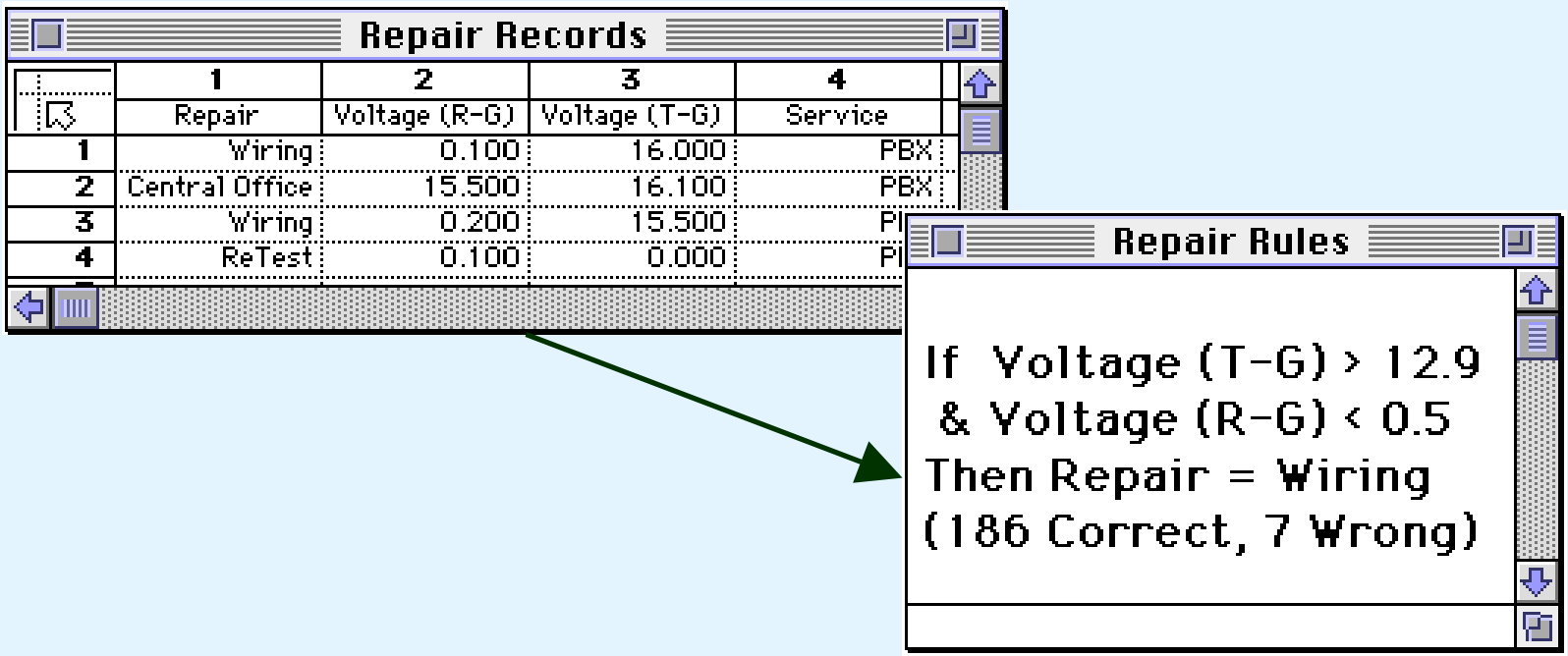
- Computational engines
- Mass storage
- Networking
- Digital libraries/data bases
- Sensors/effectors
- Software
- Services
- All organized to permit the effective and efficient building of applications


Users
App Sys
App Sys
App Sys
EFFECTORS
Cyberinfrastructure
STORAGE & LIBRARIES
CYCLES
SOFTWARE
SERVICES
SENSORS EFFECTORS
High Bandwidth Networks

From Data to Decisions: Successes

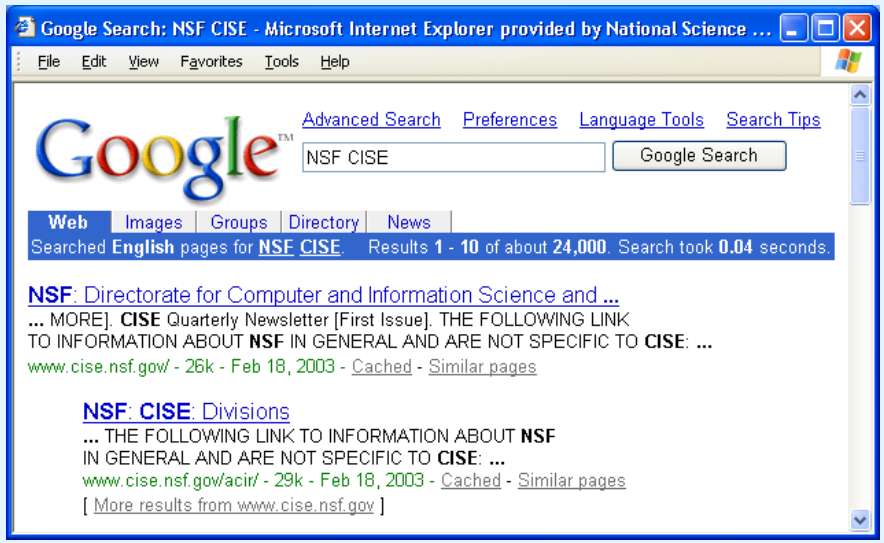
Data Mining

"… drowning in data but starving for knowledge"



Successes: Internet Search

- It's quicker to find a paper on the Internet than on your bookshelf
- There's no longer a need to remember URLs.



Challenges for Data Analysis

- Learning from any data representation, e.g., relational data, transactional data, text, images, etc.
- Total Information Awareness for scientists and engineers
  - Locating a data sources that contains information
  - Integrating information from several data sources
  - Extracting information from text, images, speech, MRIs etc.
- Explaining Discovered Knowledge in terms people understand

The NSF Cyberinfrastructure Objective

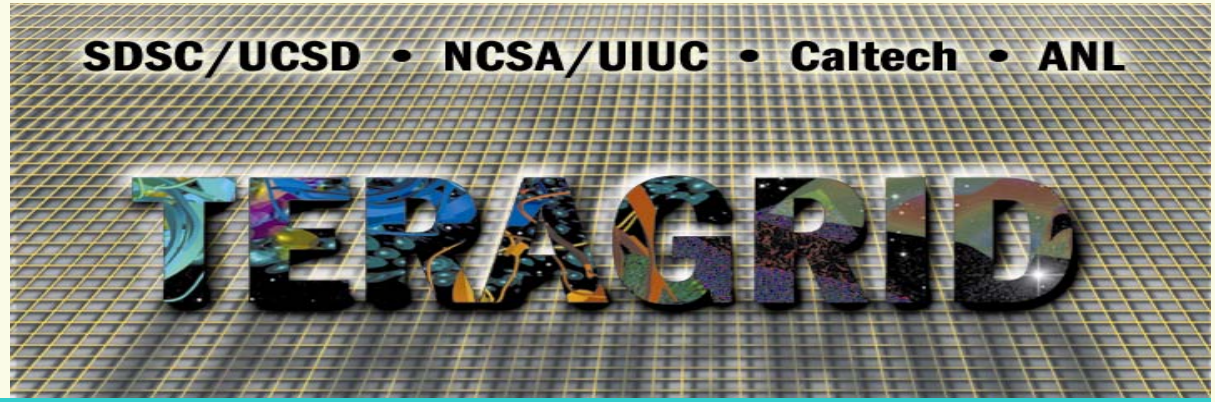
- To provide an integrated, high-end system of computing, data facilities, connectivity, software, services, and instruments that ...

- enables all scientists and engineers to work in new ways on advanced research problems that would not otherwise be solvable

The New Frontier in Grid Computing

August 9, 2001: NSF Awarded $53,000,000 to SDSC/NPACI and NCSA/Alliance for TeraGrid

October 10, 2002: NSF awarded $35,000,000 to TG team and PSC for Extended Terascale Facility



TeraGrid will provide in aggregate

- *Over 13.6 trillion calculations per second*
- *Over 600 trillion bytes of immediately accessible data*
- *40 gigabit per second network speed*

TeraGrid will provide a new paradigm for data-oriented computing

Critical for disaster response, genomics, environmental modeling, …

TeraGrid: Setting land-speed records for data

- **SC'02 Experiment: Data sent in real-time from Baltimore to San Diego and back in record time: 721 MB/sec across country**
  - ***All disk looked local*: Experiment demonstrated that data could be treated as local disk-to-disk transfer to remote processes running across TeraGrid**
  - ***Sun and SDSC collaboration on disk and software made it possible for multiple technologies to work together.***
- **828 MB/sec data transfer from disk to tape demonstrated in Fall, 2002 at SDSC**
- **Why is this important?**
  - ***Fast remote data transfer enables applications like NVO to be executed at large-scale***
  - ***Fast remote transfer makes it feasible to access whole dataset, compare multiple NVO sky surveys***

Challenges

- How to build the components?
  - Networks, processors, storage devices, sensors, software
- How to shape the technical architecture?
  - Pervasive, many cyberinfrastructures, constantly evolving/changing capabilities
- How to operate it?
- How to use it?

CS&E Research Challenges

- **Networks**: scalability, adaptivity, security, QoS, interoperability, congestion control
- **Software engineering**: verifiability of results, automated specification and generation of code, complex system design
- **Distributed systems**: theoretical foundations, new architectures, interoperability, resource management
- **High-performance computing**: new processor design, inter-process communication, performance

CS&E Research Challenges (cont.)

- **Middleware**: basic CI operation, sensor networks, data manipulation, disciplinary tools
- **Theory & algorithms**: verifiability of access and information flows, authentication, performance analysis, algorithm design
- **Sensing & signal processing**: distributed signal processing, classical problems under new constraints, fusion
- **Visualization & information management**: new models for massive datasets, resource sharing, knowledge discovery

Conclusion

- Cyberinfrastructure in some ways is just the natural next stage of computer usage
- It differs in the ubiquity, interconnectedness, and power of available resources
- It thus is engendering a revolution in S&E
- It is critical to the advancement of all areas of S&E
- It provides a plethora of interesting and deep research challenges for CS&E

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

Dr. Peter A. Freeman
NSF Assistant Director for CISE
Phone: 703-292-8900
Email: pfreeman@nsf.gov

Visit the NSF Web site at:
www.nsf.gov