# Progress and Challenges in Informatics: A Continuing Explosion of Data

**Kathie L. Olsen, Ph.D.**

**Associate Director for Science**

**Office of Science and Technology Policy**

**April 26, 2004**

Photo credit: New Granule Neurons
Heather A. Cameron, Ph.D.,NIMH Investigator

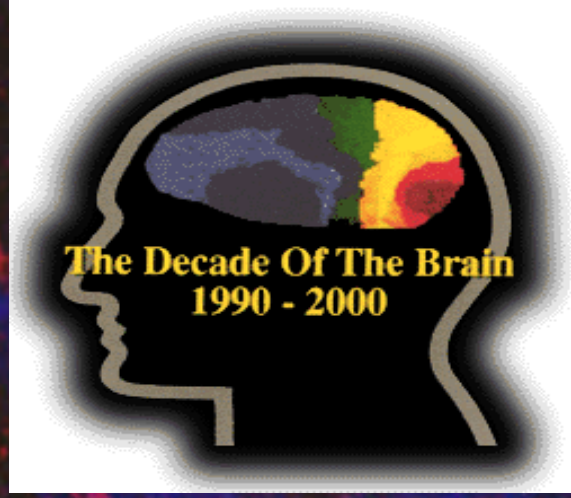
# Informatics for Neuroscience

- How did we get to today?
- How do we capitalize on advances in other disciples?
- How do we expand internationally?
- How do we take our vision of integrative networks to capture, manage, store, & utilize neural data in the 21st Century?
- How to enable access by others than those originally obtained the data?

# Role of White House Office of Science & Technology Policy

- Advise the President & Offices of the President
- Lead the interagency effort to develop S&T policies and budgets for ALL AREAS OF SCIENCE
  - Co-chair National Science & Technology Council (NSTC)
- Build strong partnerships among federal, state and local governments, OTHER COUNTRIES, industry, academia & scientific associations
- Develop clear, measurable goals and objectives for R&D programs
- Access Federal investments relative to purposes of government

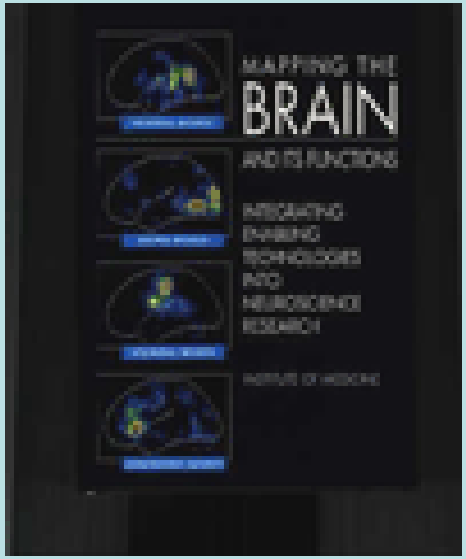
# Decade of the Brain




- President George Bush designated the 1990's as the **Decade of the Brain**
  - Subcommittee on Brain & Behavioral Sciences at OSTP was formed
    - Roger Porter (NINDS)-Chair
    - Alan Leshner (ADAMHA) & Steve Zornetzer (ONR)-Vice Chairs
    - Maximizing Human Potential Published
  - From 1990 to the end of 1999, the Library of Congress and NIMH sponsored a unique interagency initiative to advance the goals set forth by "to enhance public awareness of the benefits to be derived from brain research"
- -

# 1991: Mapping the Brain

- NAS study "Mapping the Brain and its Functions: Integrating Enabling Technologies into Neuroscience Research"
  - Study funded by NSF, NIH
  - Methods for collecting and integrating the current "explosion of data"
  - Progress requires new capabilities in computers and information science



# NIMH: The Human Brain Project

# NSF: Computational Neuroscience Program

- Research programs resulting from recommendations in NAS report
- Support research and development of advanced technologies, and infrastructure support
- Cooperative efforts among neuroscientists and information scientists

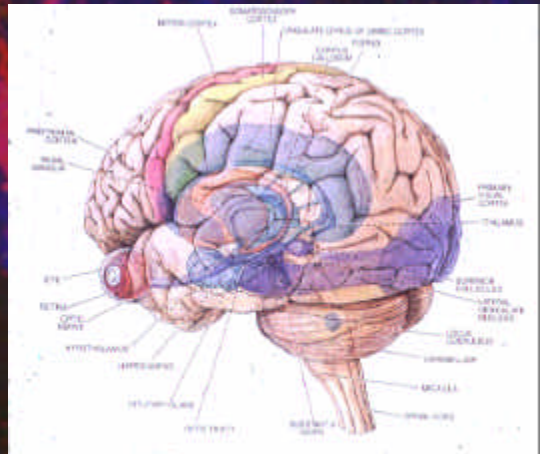
Neuroinformatics

The Human Brain Project

# Informatics for Neuroscience

## What's needed?

- Science/User requirements
- Technology
- Partnerships
- Infrastructure
- Training, Education, Capacity development
- Policy
- Accountability



# Science/User Requirements

- Integration and interoperability for disparate information sources
  - Large amounts
  - Different formats
  - Raw data, metadata, and derived products
- Analysis
- Spatial and temporal visualization
- Modeling
- Archival and retrieval
- Data continuity, validation, verification
- User requirements that are ALWAYS ahead of the technology development

# Evolving Science Requirements for Networks

| Science Areas | Today *End2End* Throughput | 5 years End2End Throughput | 5-10 Years End2End Throughput | Remarks |
|---|---|---|---|---|
| High Energy Physics | 0.5 Gb/s | 100 Gb/s | 1000 Gb/s | High bulk throughput |
| Climate (Data & Computation) | 0.5 Gb/s | 160-200 Gb/s | N x 1000 Gb/s | High bulk throughput |
| SNS NanoScience Spallation Neutron Source | Not yet started | 1 Gb/s | 1000 Gb/s + QoS for Control Channel | Remote control and time critical throughput |
| Fusion Energy | 0.066 Gb/s (500 MB/s burst) | 0.198 Gb/s (500MB/ 20 sec. burst) | N x 1000 Gb/s | Time critical throughput |
| Astrophysics | 0.013 Gb/s (1 TByte/week) | NN multicast | 1000 Gb/s | Computational steering and collaborations |
| Genomics Data & Computation | 0.091 Gb/s (1 TBy/day) | 100s of users | 1000 Gb/s + QoS for Control Channel | High throughput and steering |

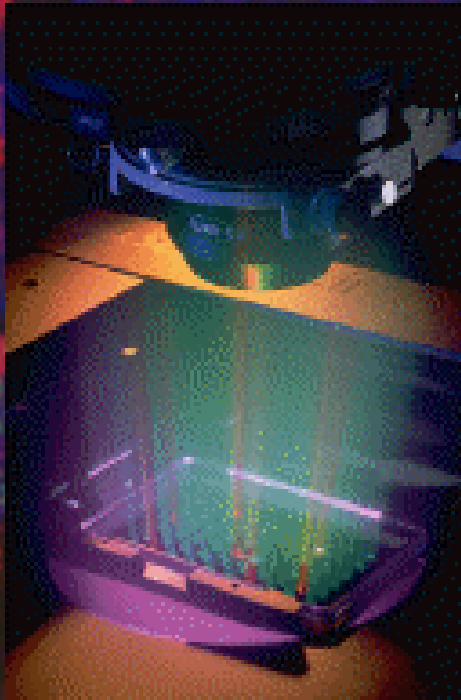
Courtesy of Harvey Newman. Caltech

# Learning from Earth Science

- Global Observations Requirements
  - Multiple end users with varying data requirements
  - Data from satellite, aircraft, in-situ
  - Varying spatial and temporal resolution
  - Multiple formats
  - Long-term data continuity and archival

# Technology

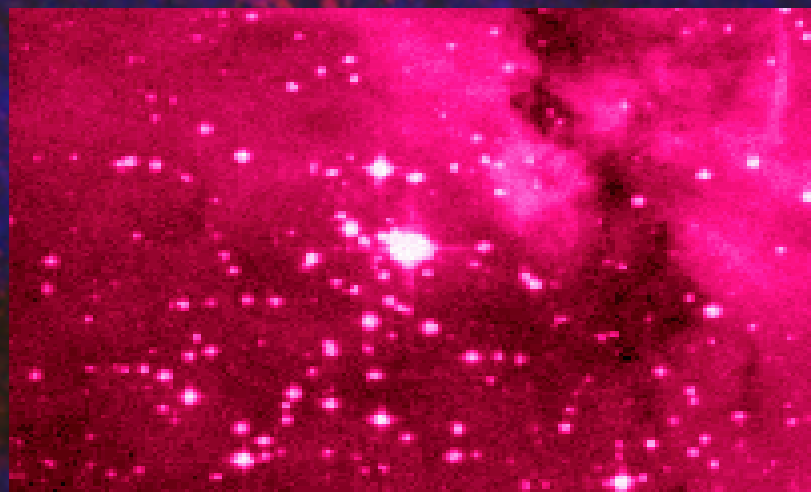
- Legacy systems
- Proliferation of tools
- Increasing complexity of tools
- Multiple hardware platforms
- Large bandwidth requirements
- Explosion of data
- Security: physical and information



# Learning from Astronomy and Physics

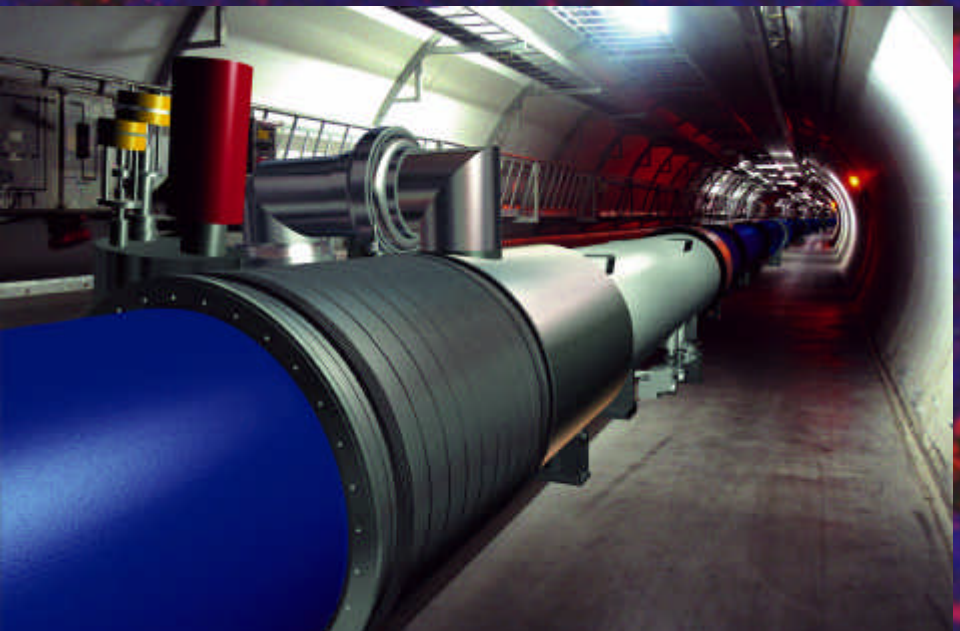
- Large area digital sky surveys
  - Combine teraflop computing requirements with terabyte storage needs
  - Volume of records is orders of magnitude larger than the largest databases manageable with today's technology
- National Virtual Observatory—Prototypes, Standards
- Radio, millimeter, infrared, optical, X-ray, gamma ray—ground-space
  - Unique detector technologies, data types, data formats, data volume,

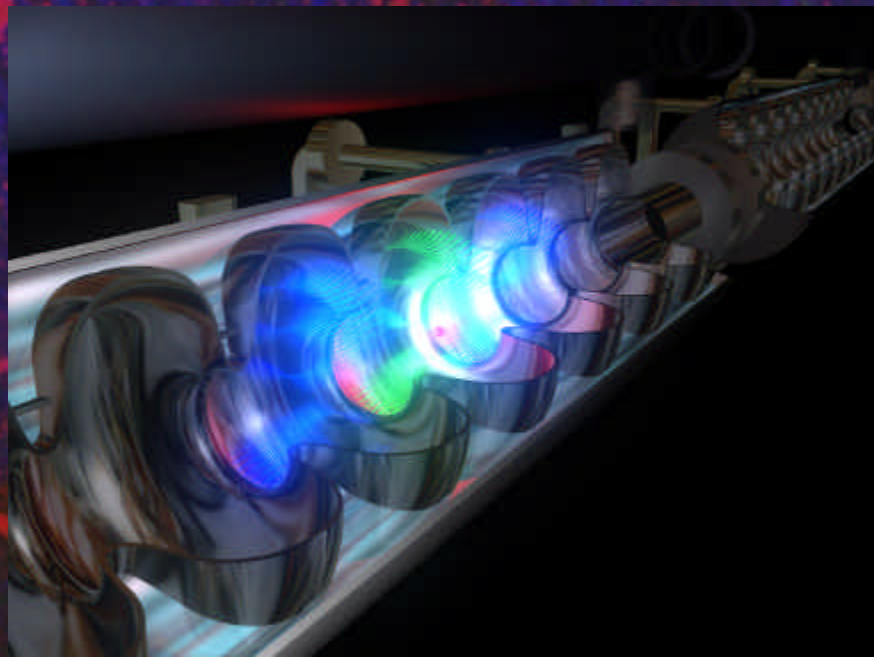
SHARING



# Learning from Astronomy and Physics

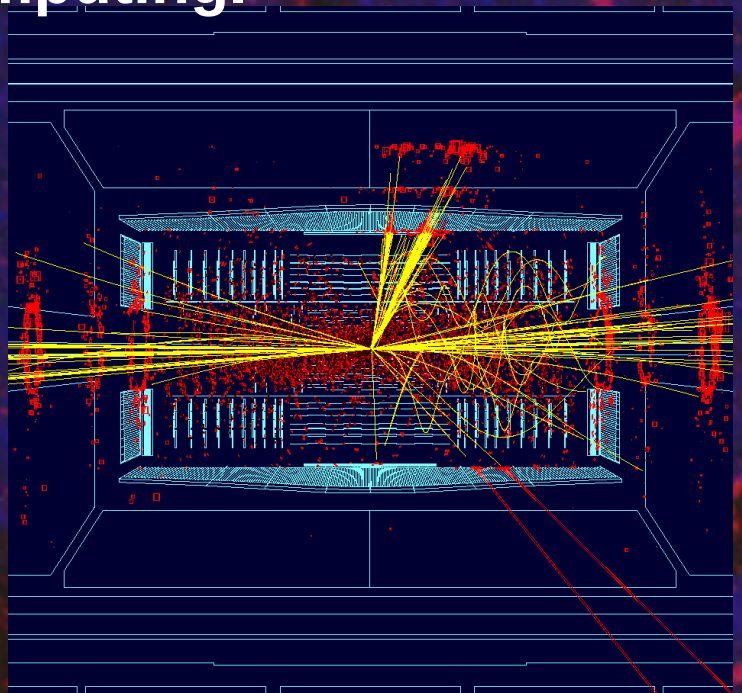
- Particle physicists have been at the vanguard of data-handling and processing technology. From the use of direct writing of data to tape in the 1970s through the development at CERN of the WORLD WIDE WEB in 1990.
- The next generation of particle physics experiments at the Large Hadron Collider (LHC) at CERN will have incredible demands.





# Learning from Astronomy and Physics

- Data from the LHC will be pre-analyzed to reduce it by a factor of more than ten million, leaving *only* a Petabyte (1015 bytes = a billion megabytes) of data per year for scientific analysis!
- The task is finding a single rare event in these data will be daunting and requires new technologies: GRID Computing.



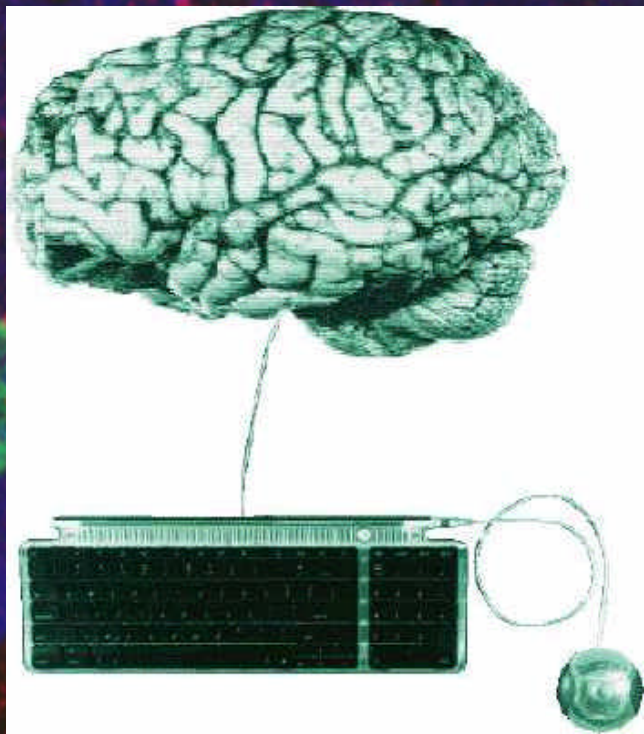
# Partnerships

- Agreements between various entities
  - Public and private partners
  - Countries and international organizations
- Licensing and intellectual property rights
  - Competing interests between intellectual property and “full and open access”
  - Purchase or license?
- Communication
- Culture, management, and economics
  - Funding issues

# Global Science Forum

- Formation of neuroinformatics subgroup in 1996—Led by Dr. Koslow
- Report in 1998
  - Establish a global informatics network
- Neuroinformatics Working Group (2000-2002) proposal:
  - National neuroinformatics research programs
  - An International Neuroinformatics Coordinating Facility
  - An international coordinated funding scheme

www.neuroinf.org
International neuroinformatics Information portal



# Global Science Forum

- International drafting committee for the International Neuroinformatics Coordinating Facility (INCF) and Program in Neuroinformatics (PIN) met April 1-2, 2004 at the OECD
- Near final versions of INCF
  - Business plan
  - Understanding
  - RFP for hosting the International Neuroinformatic Coordinating Facility
- PIN (funding program) decoupled from INCF
- Final meeting in June
  - following that meeting the Understanding will be given to ministers for their approval
  - RFP to be released and with the goal of having a host site and the opening of the INCF in early 2005.

# Partnerships

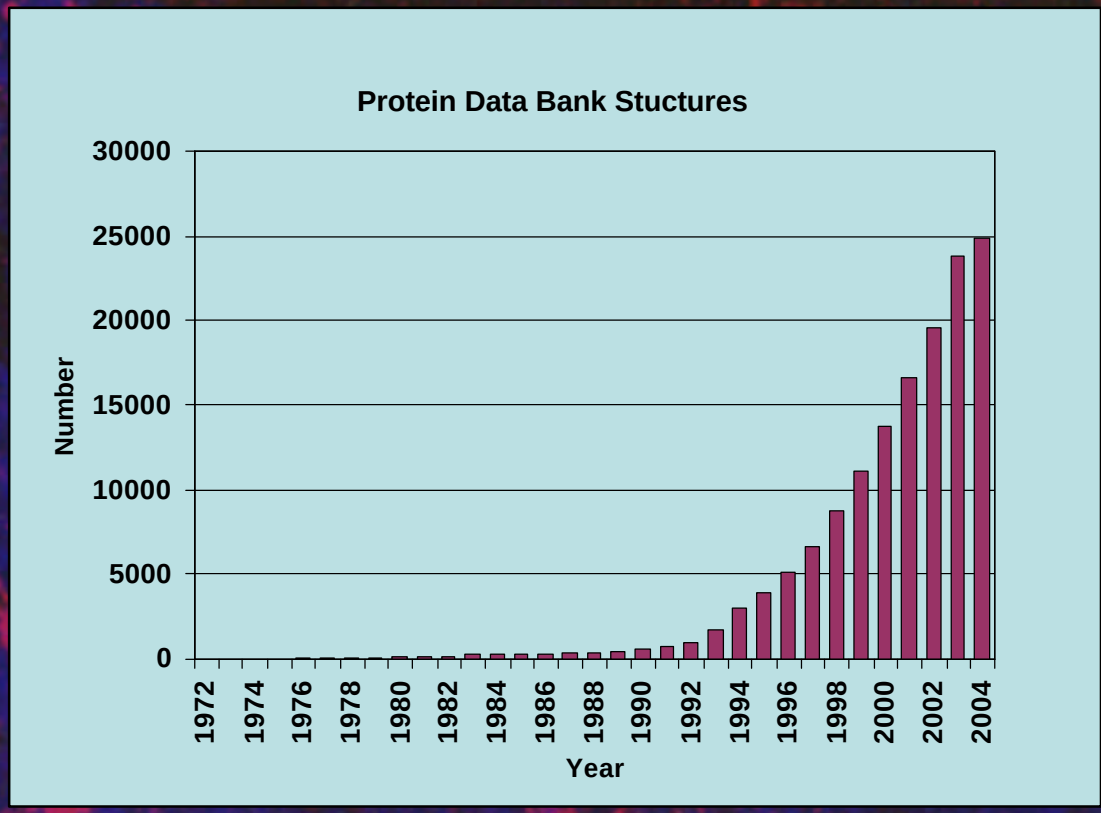
- Different types of people working together
  - Brain scientists
  - Physicists, chemists, social scientists, etc.
  - Computer engineers/innovators
  - Mathematicians/software engineers
  - Students

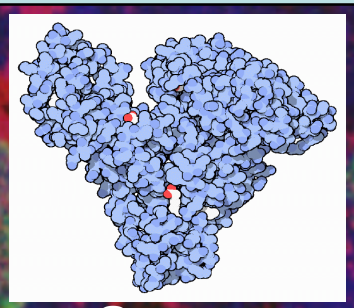
# Infrastructure

- Difficulty of inserting advanced technologies into existing systems
- Difficulty of integrating new initiatives with other existing initiatives
- Funding for new
  - Physical
  - IT
  - Data services
- Maintenance and upgrades for existing
- Long term support issues
  - Archival
  - Data conversion to new formats

# EXAMPLE: Protein Data Bank

- Community discussions regarding need started in the 1960s.
- Established at Brookhaven National Lab in 1971 with 1 structure, grown to 25,000 in 2004.
- A rapid expansion of PDB in the 1980s with the advent of synchrotron light source, other technological developments for structural biology.
- Deposits also increased as funding agents required "open" access to data.
- Evolution of technology: punch cards, magnetic tape, now electronically.



Serum Albumin

Quality of data
Credit for deposits
Sharing

# Training, Education, and Capacity Development

- Increased access
- Optimal utility
- Develop the next generation
- NSTC Committee on Science-Subcommittee on Research Business Models:
  - New Ways of doing Research

    Funding Mechanisms

    Rewards Systems

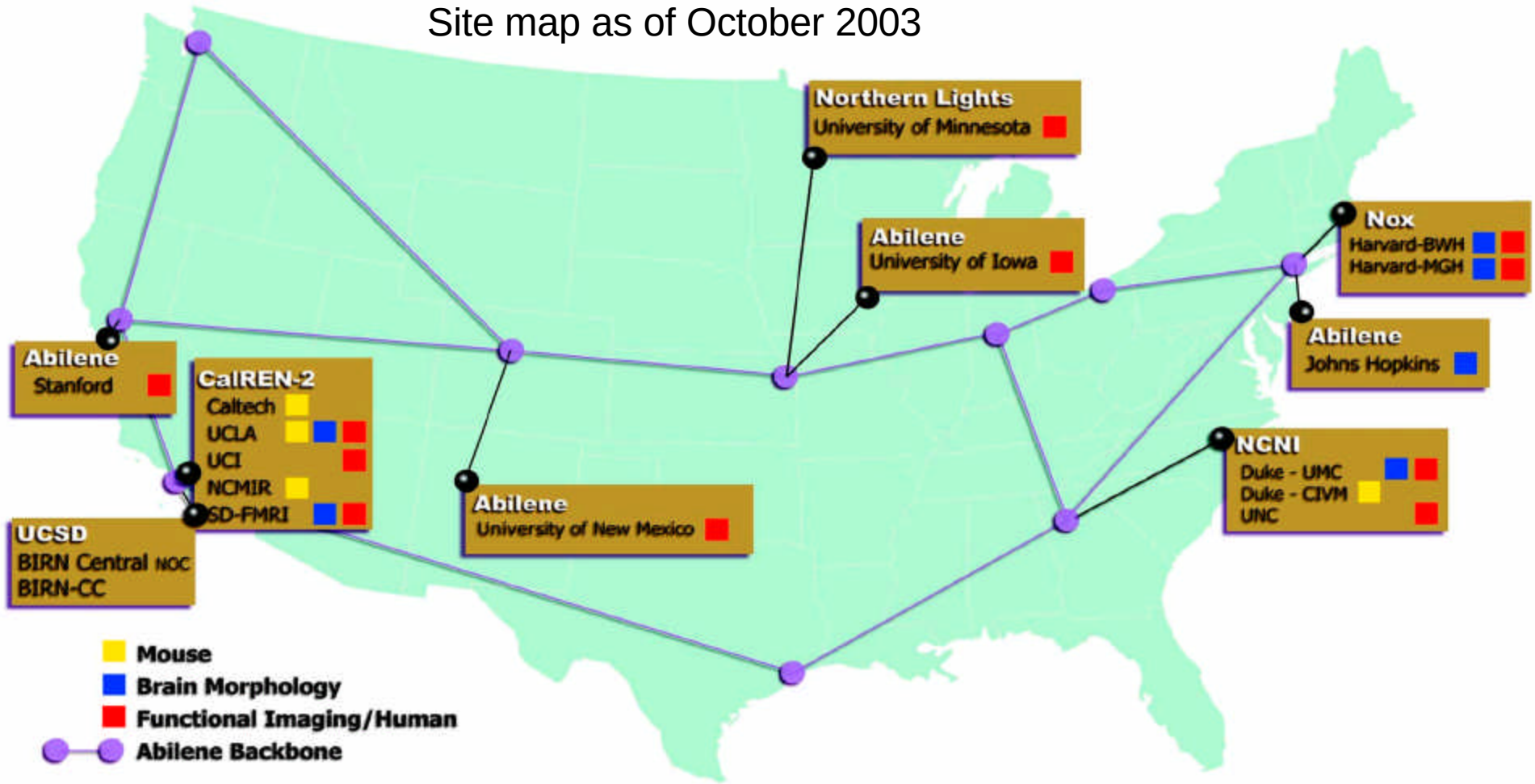
    Training Opportunities

(Connie Atwell, Nat Pitts)

Student exploring 3D visual environment

# Biomedical Informatics Research Network (BIRN)

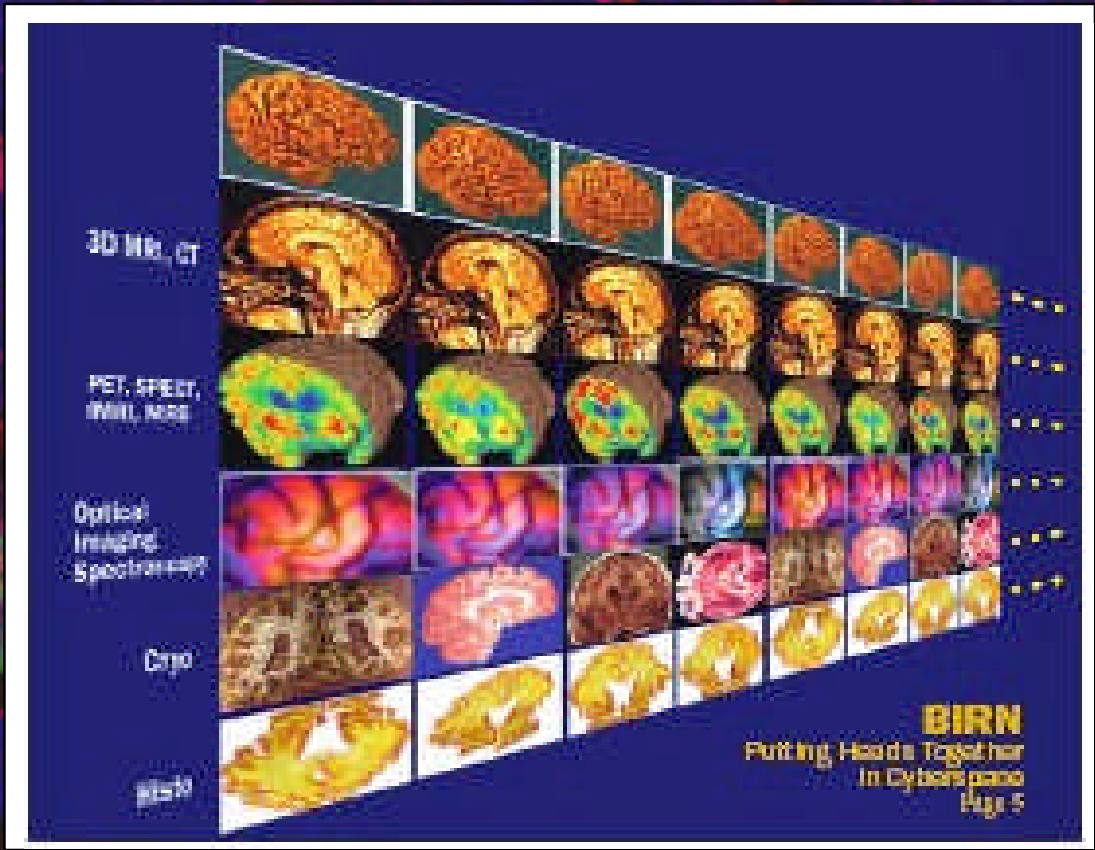
Site map as of October 2003




A shared biomedical IT Infrastructure to hasten the derivation of new understanding and treatment of disease through use of distributed knowledge

# What is BIRN?

- Develop solutions to large scale biomedical data sharing and mining and the need to access imaging tools
- Testbeds and a coordinating center
- Create algorithms and other tools that are immediately useful but at the same time are extensible and scalable to other disciplines

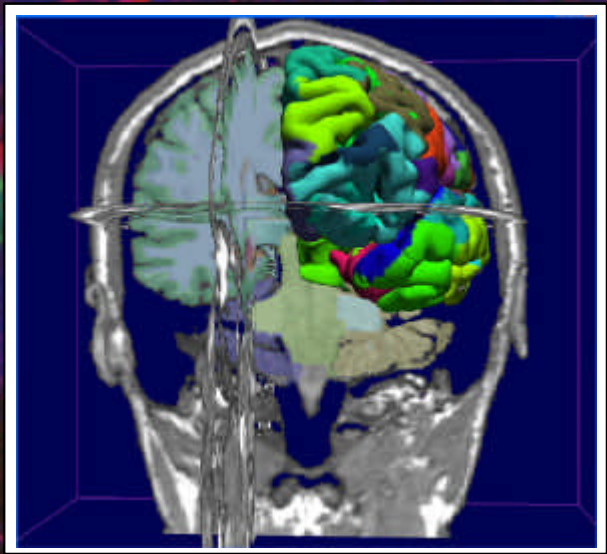


# Policy

- Roles and responsibilities
  - International agreements
- Address competing priorities
  - Obtaining data
  - Creation of products
  - New capabilities
- Standards
- Data quality, assurance, availability, and access
- Planning and Advice
  - Governing bodies
  - Advisory committees
  - Working groups

# Accountability

- Performance Measures
  - Meeting the needs of the various users?
  - Usage and science output
  - Return on investment for industry
- Cost Accounting
  - Cost-effective performance
  - Validation and verification of costs
- Determining the pay-off
- Articulating the pay-off



The Decade Of The Brain
1990 - 2000
YOU ARE HERE

# “The Universe is Boundless and the Discoveries Unlimited”

E. Weiler, Ph.D.

Associate Administrator for Space Science, NASA