“Information Infrastructure for the Social Sciences in the 21st Century”

Hubert M. Blalock, Jr. Memorial Lecture Series on Advanced Topics in Social Research

University of Michigan

July 13,1998

Dr. Larry Smarr

Director, National Computational Science Alliance and the National Center for Supercomputing Applications

Professor in the Departments of Physics and Astronomy

University of Illinois Urbana-Champaign

Emerging Computational Trends and the Quantitative Social Sciences

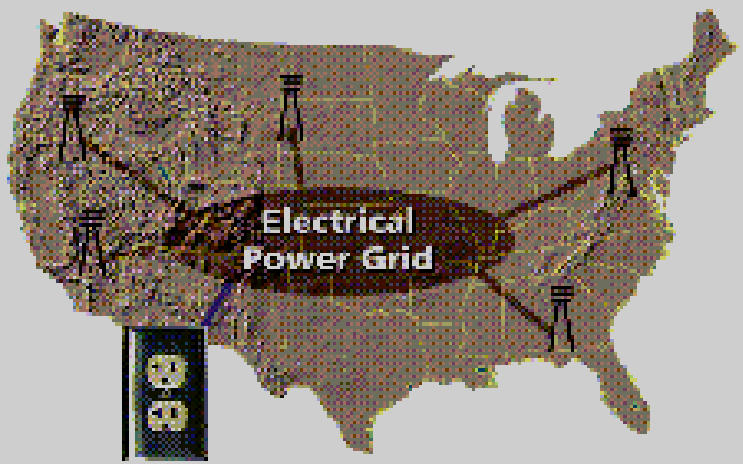
- Grid Technologies
- Document and Data Management
- Information Visualization
- Web Computing
- Scalable Computing

Behavioral and Social Sciences in the 21st Century
Philip Smith and Barbara Torrey

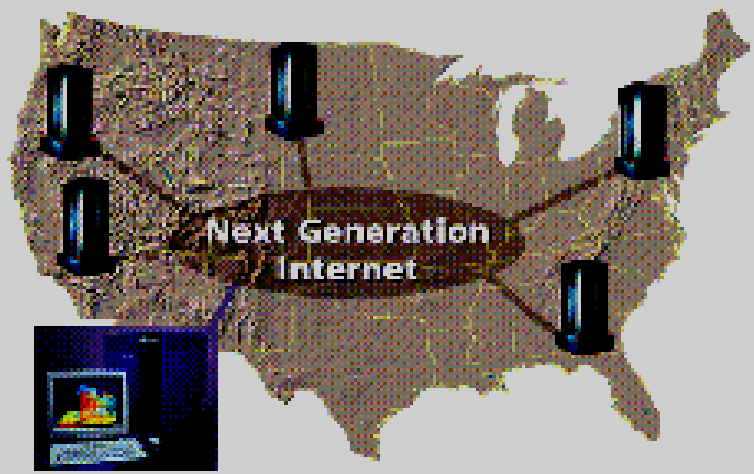
- Integrate Current Data Sets
- Improve the Coverage of Longitudinal Studies
- Experiment with Nonlinear Dynamic Systems
- Develop Comparable International Research
- Integrate Quantitative and Qualitative Research to Advance New Theory

The Emerging Concept of a National Scale Information Power Grid

Imagine a national computing and information infrastructure that allowed everyone to access the information resources of the nation in much the same way that one accesses electrical power today; an "Information Power Grid"

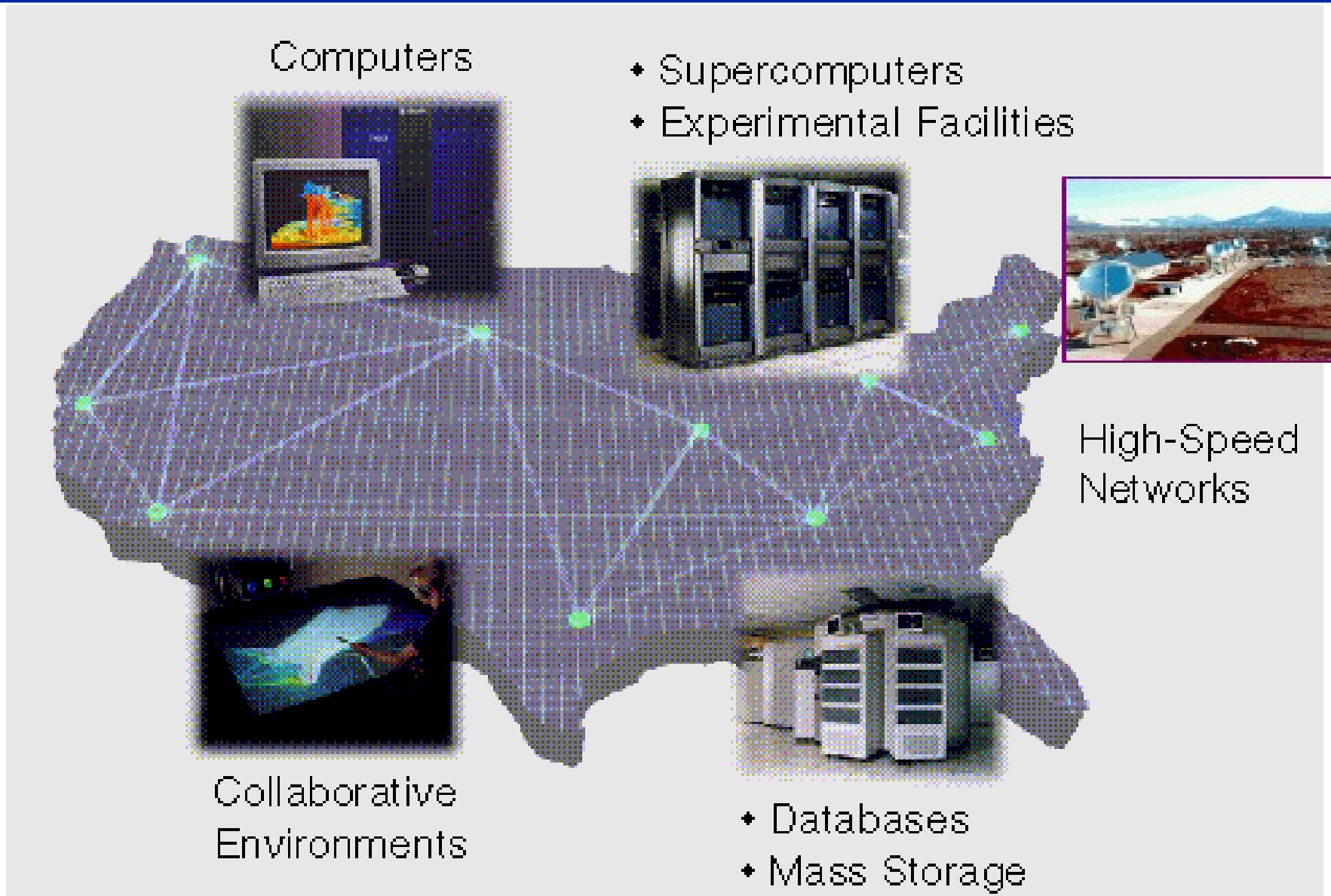


Your Electrical Power
Available Here



Your Information Power
Available Here

The Grid Links People with Distributed Resources on a National Scale



The National Center for Supercomputing Applications

- Is a Federal / State / University / Industry Funded Center
  - Budget $50 Million/Year
  - 500 Work at NCSA
- Is a Unit of the University of Illinois at Urbana-Champaign
- Has a Mission of Providing Access to Leading Edge Information Technologies to Universities and Industry
- Had Major Influence on the Creation of:
  - The Internet
  - The Web
  - Scientific Visualization
  - Computational Science, Engineering, and Knowledge Management

NCSA is the Leading Edge Site for the National Computational Science Alliance



Alliance National Technology Grid

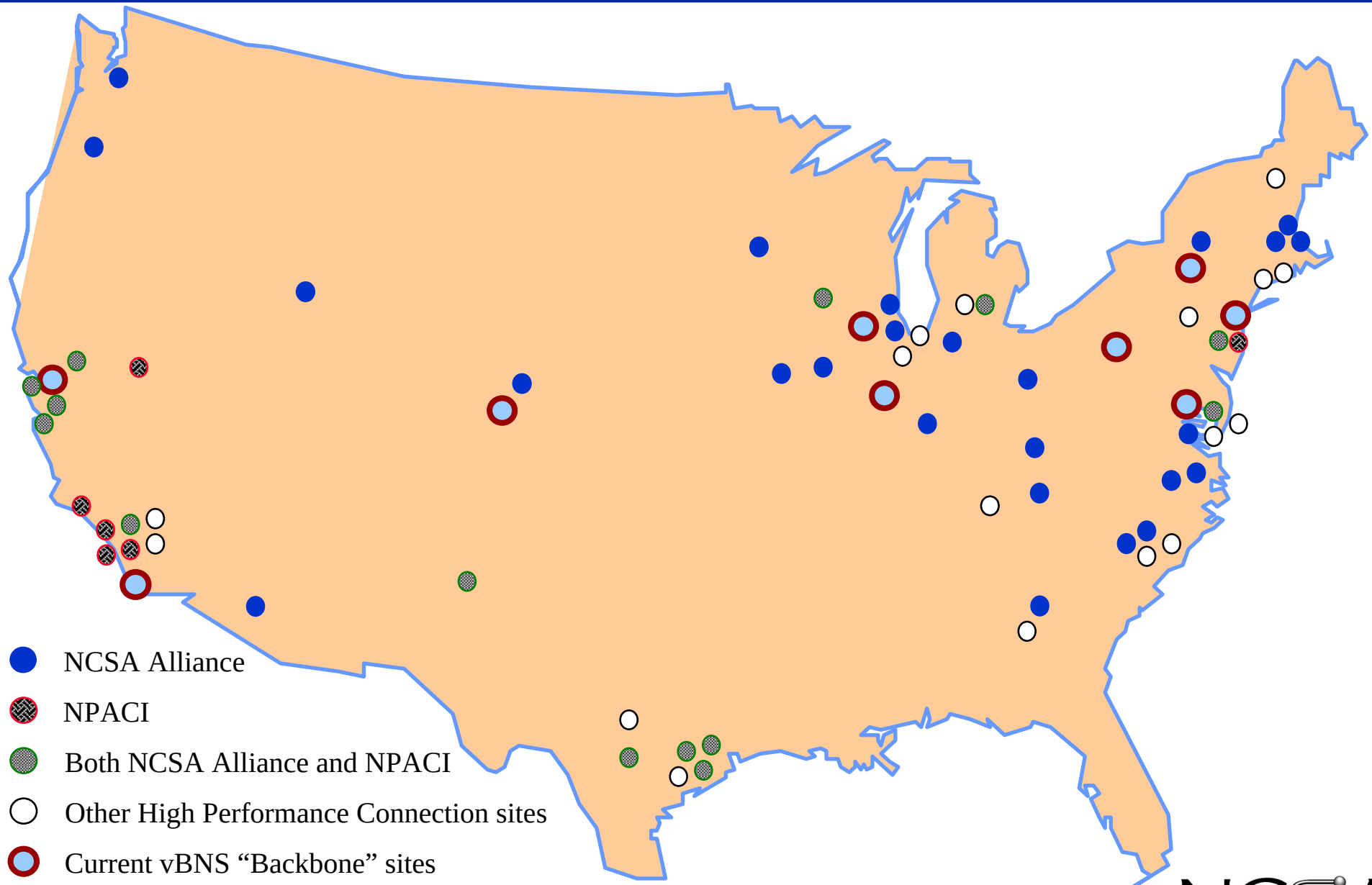
www.ncsa.uiuc.edu

NCSA™
National Computational Science Alliance

The Alliance Team Structure to Prototype the 21st Century Information Infrastructure

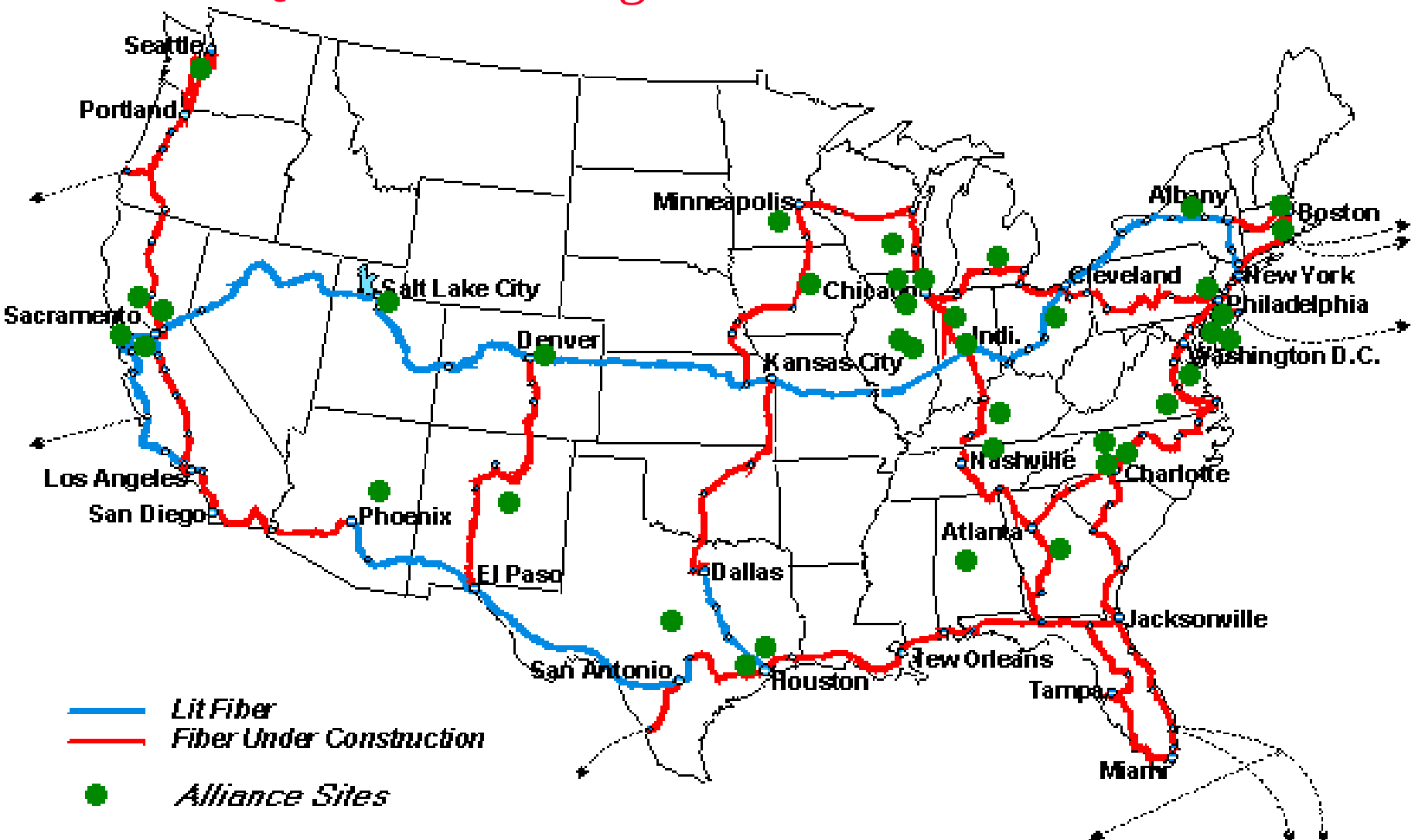
- **Leading Edge Center**
- **Enabling Technology**
  - Parallel Computing
  - Distributed Computing
  - Data and Collab. Computing
- **Partners for Advanced Computational Services**
  - Communities
  - Training
  - Technology Deployment
  - Comp. Resources & Services
- **Strategic Industrial and Technology Partners**
- **Application Technologies**
  - Cosmology
  - Environmental Hydrology
  - Chemical Engineering
  - Nanomaterials
  - Bioinformatics
  - Scientific Instruments
- **EOT**
  - Education
  - Evaluation
  - Universal Access
  - Government

NSF vBNS and PACI - Mutually Interdependent



FY99 Qwest Nationwide Network - Backbone for Internet2 Abilene - More Links

Qwest Partnering with Cisco and Nortel



http://www.qwest.net/network/Mainmaps.html

Source: Randy Butler, NCSA

NCSA™
National Computational Science Alliance

Alliance National Technology Grid Workshop and Training Facilities

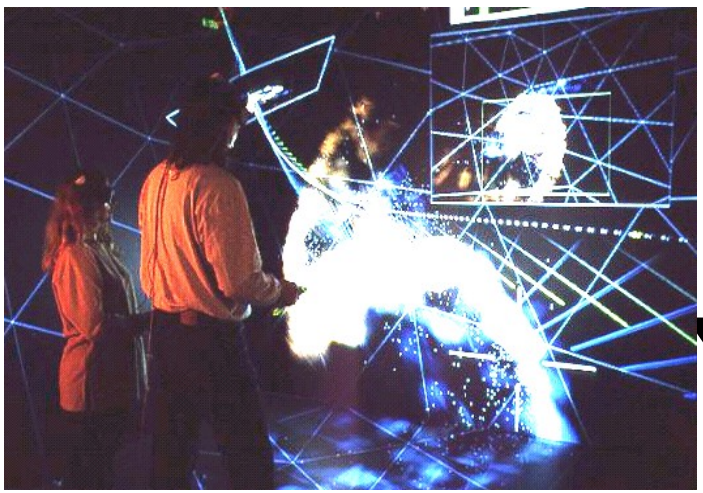


Being Deployed Across the Alliance

Jason Leigh and Tom DeFanti, EVL; Rick Stevens, ANL

Integrating Digital Video With the Grid

Interactive Virtual Environments



Application Teams

Internet, vBNS

Desktop Video Conferencing

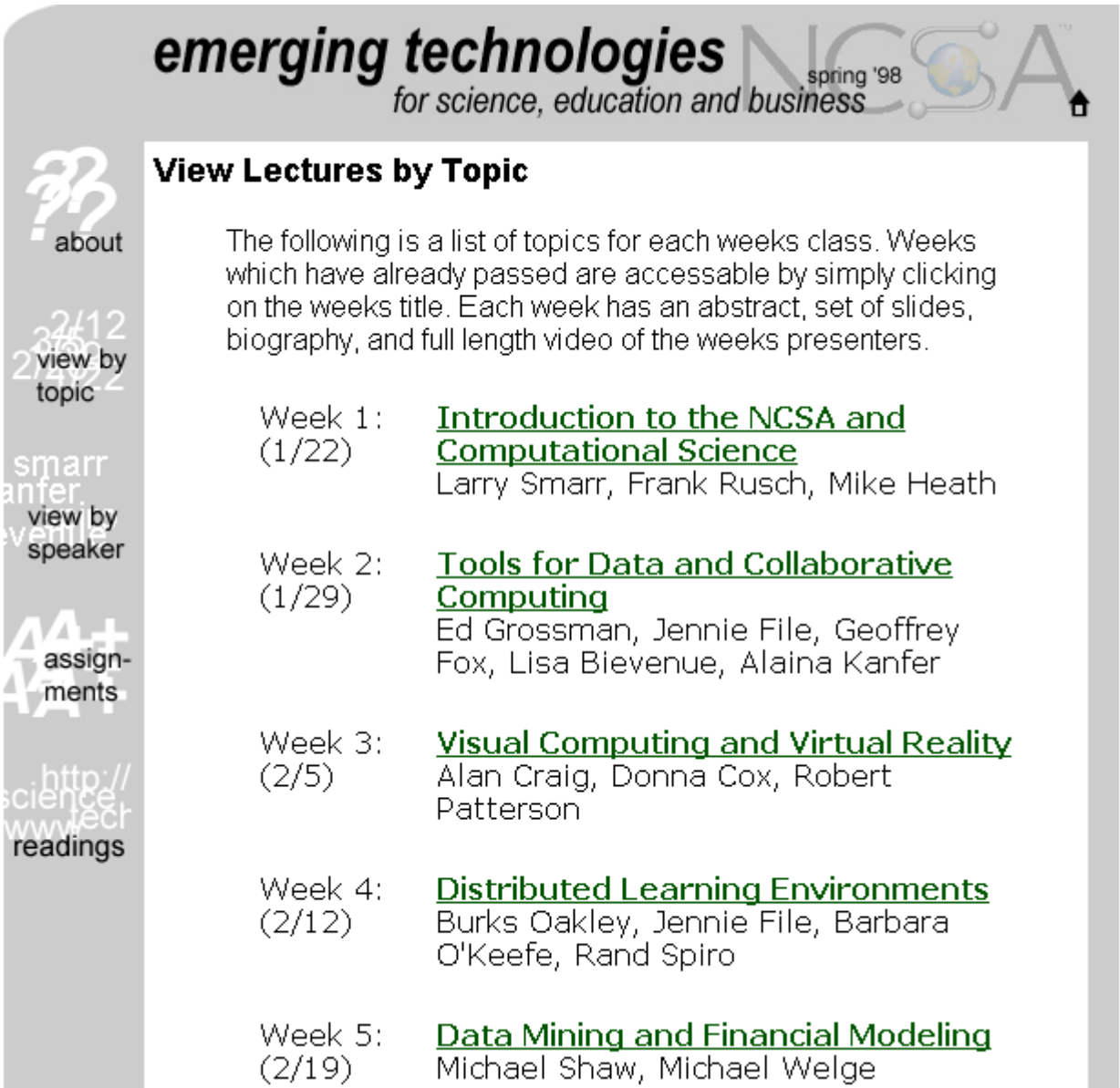


Digital Video Server

Individual Desktops

Create Digital Video Animation Concurrently with Supercomputing

Alliance Emerging Technologies Course on Streaming Video



- NCSA has 20 courses
- Alliance Goal of 100 by end of 1998

Alliance'98 Talks Were Webcast and Archived

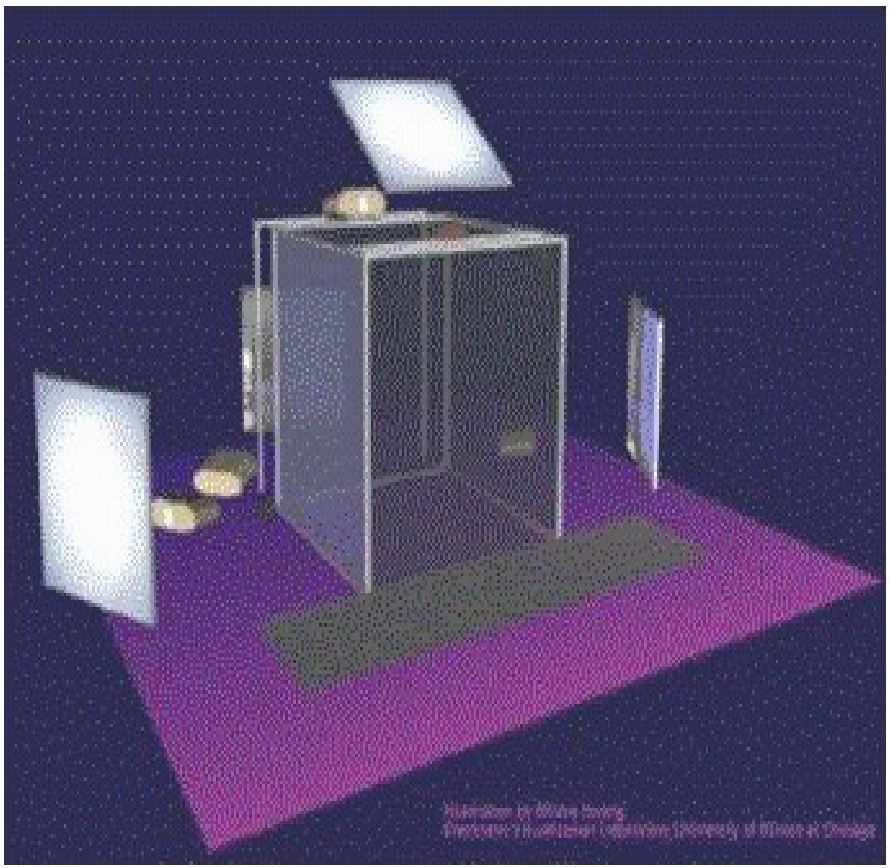
http://www.ncsa.uiuc.edu/edu/course98/lecturers/week/

High Performance Geographic Information Systems

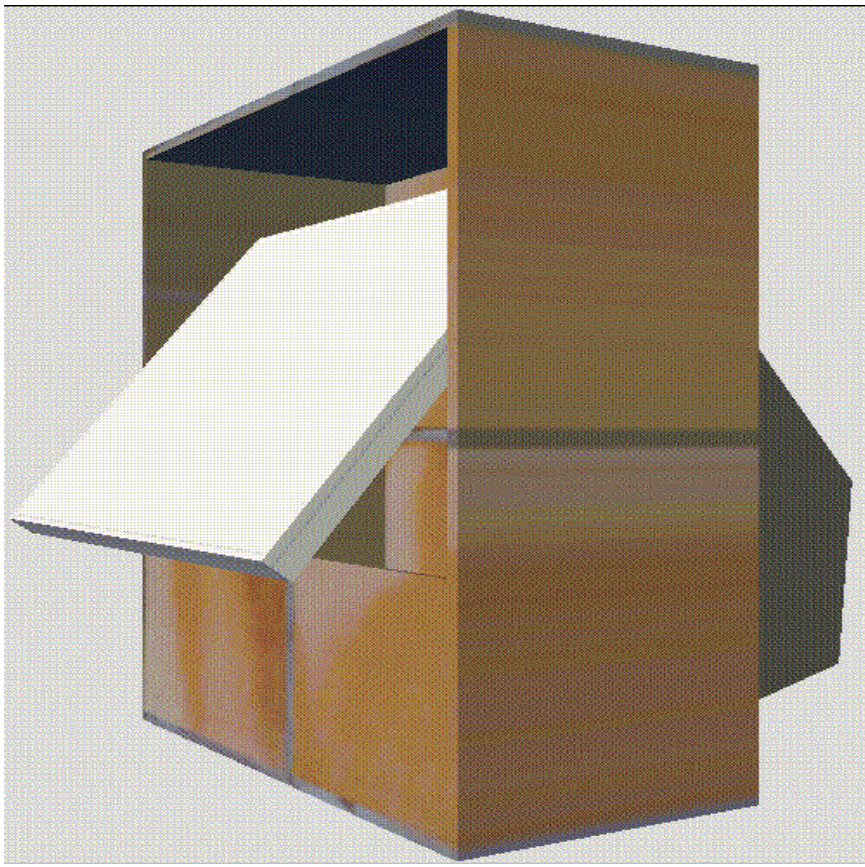
- **HPGIS (NCSA)**
  - Large Datasets Spatially or Temporally
  - Use of CAVE to Render GIS Objects
  - Parallel Computing and I/O
  - Collaborative Interactive Investigations
- **Drivers**
  - NSF PACI-Environmental Hydrology
  - Digital Government (Federal Application Council)
  - Digital Earth (Gore)
  - NASA / Mission to Planet Earth
  - DOE Strategic Simulation Program-Global Change

Source: Doug Johnston, NCSA, UIUC

The Killer Application for the Grid - Collaborative Tele-Immersion



CAVE

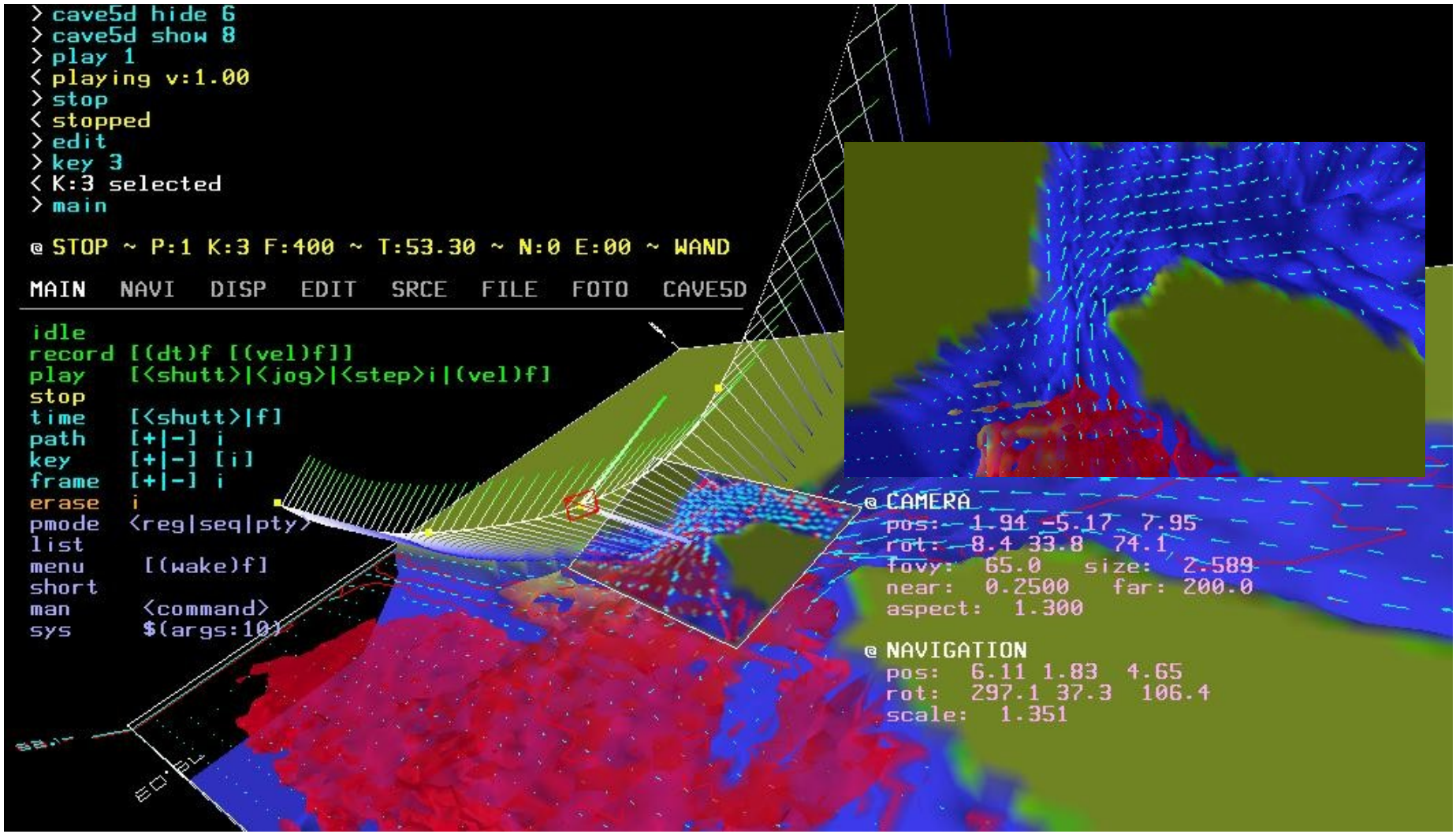


ImmersaDesk

Different Physical Implementations of the Alliance CAVE Software Libraries

Image courtesy: Electronic Visualization Laboratory, UIUC

Goal-Analyze and Record Complex Data sets Using Interactive Virtual Environments

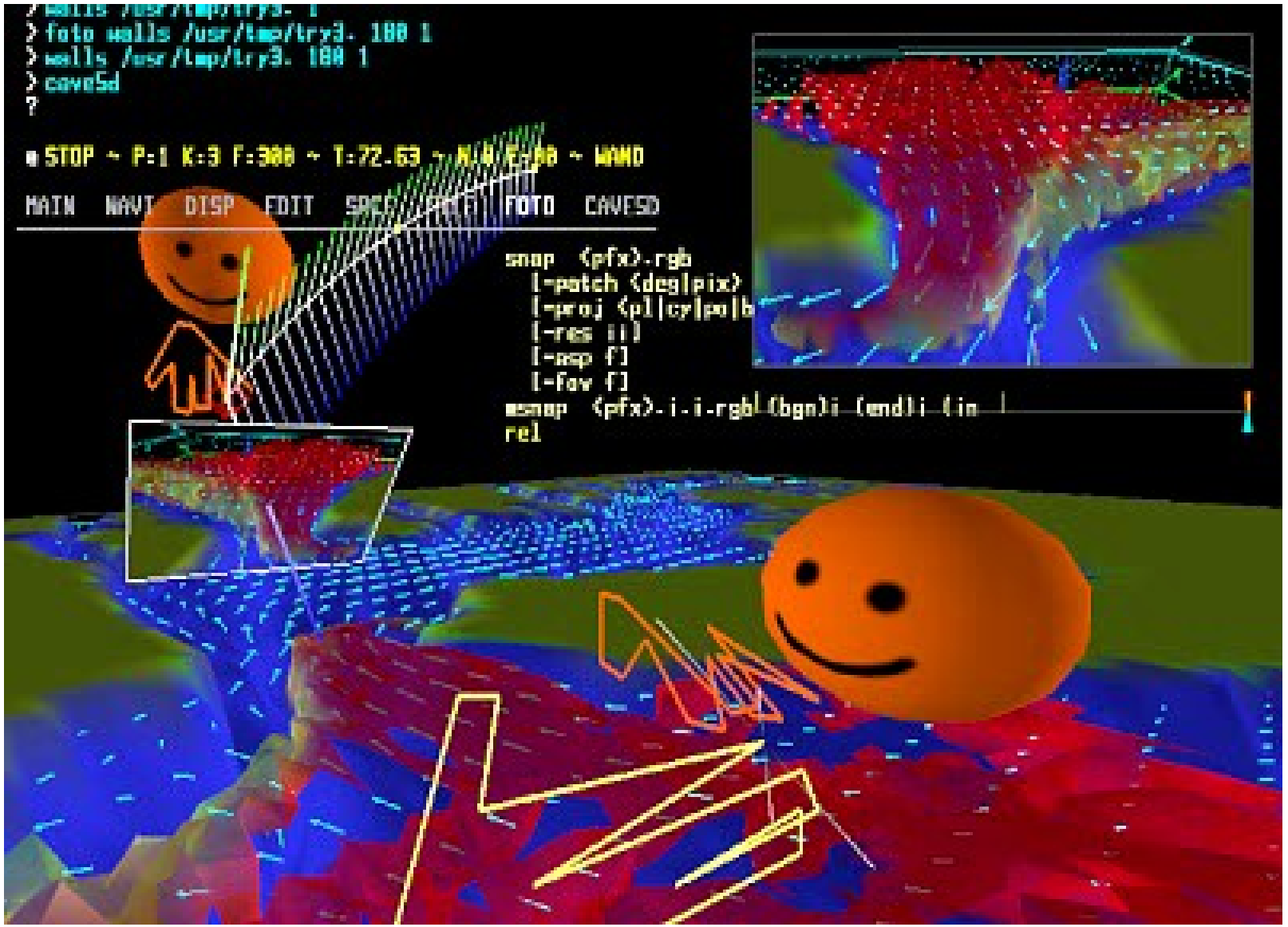


Cave5d Enables Interactive Visualizations of Time-Varying, 3-Dimensional Vis5d Data Sets in CAVE Environments

Donna Cox, Robert Patterson, Stuart Levy, NCSAVirtual Director Team
Glenn Wheless, Cathy Lascara, Old Dominion Univ.

National Computational Science Alliance

Avatars Show Head & Hand Pointing in Shared Virtual Space



Donna Cox, Robert Patterson, Stuart Levy,
NCSAVirtual Director Team

Goal-Create Shared Virtual Environment

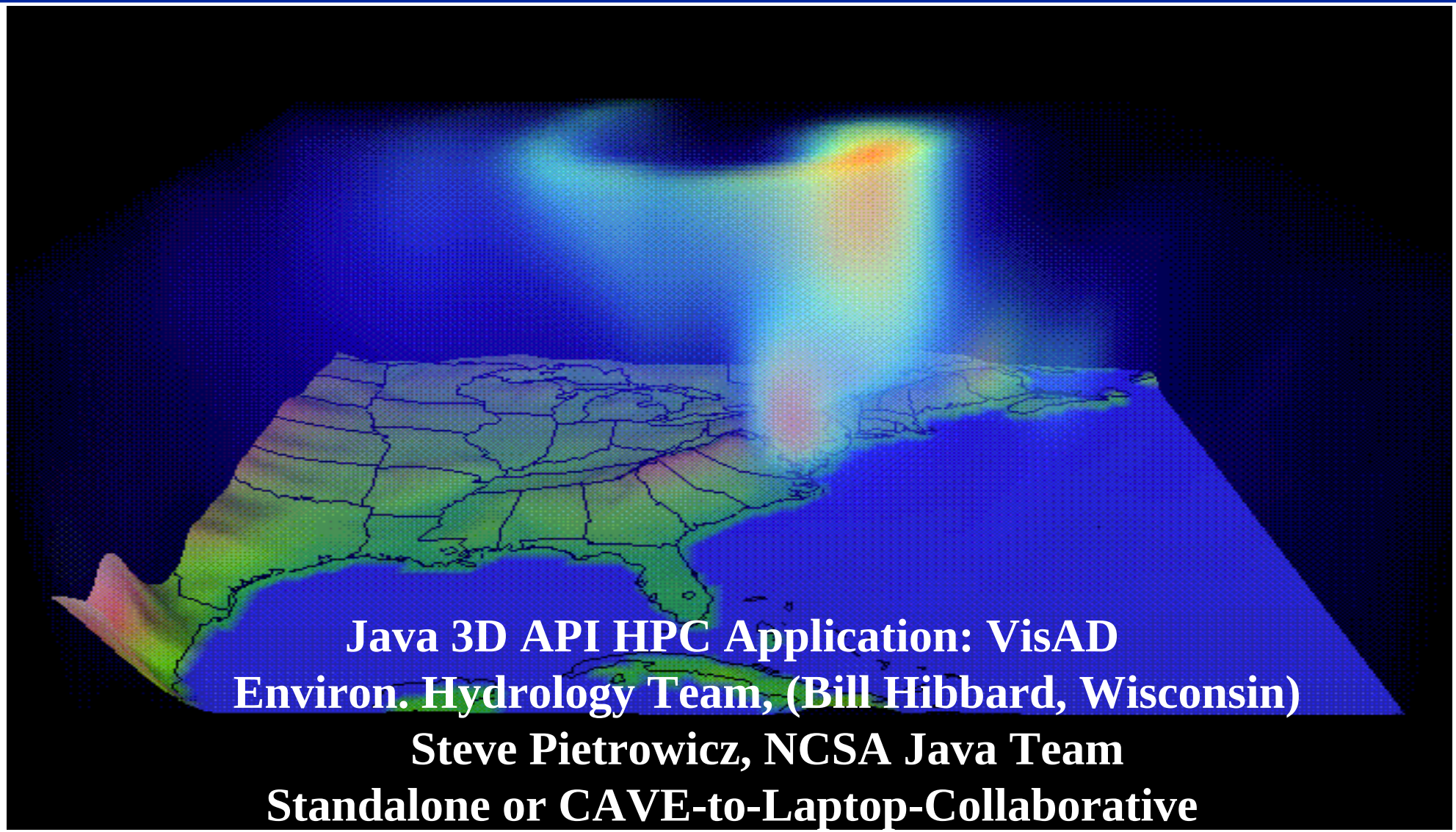
CVD -- Collaborative Virtual Director



Donna Cox, Robert Patterson, Stuart Levy, NCSAVirtual Director Team
Glenn Wheless, Old Dominion Univ.

NCSA™
National Computational Science Alliance

Goal-Linking the CAVE to the Desktop: Collaborative Java3D



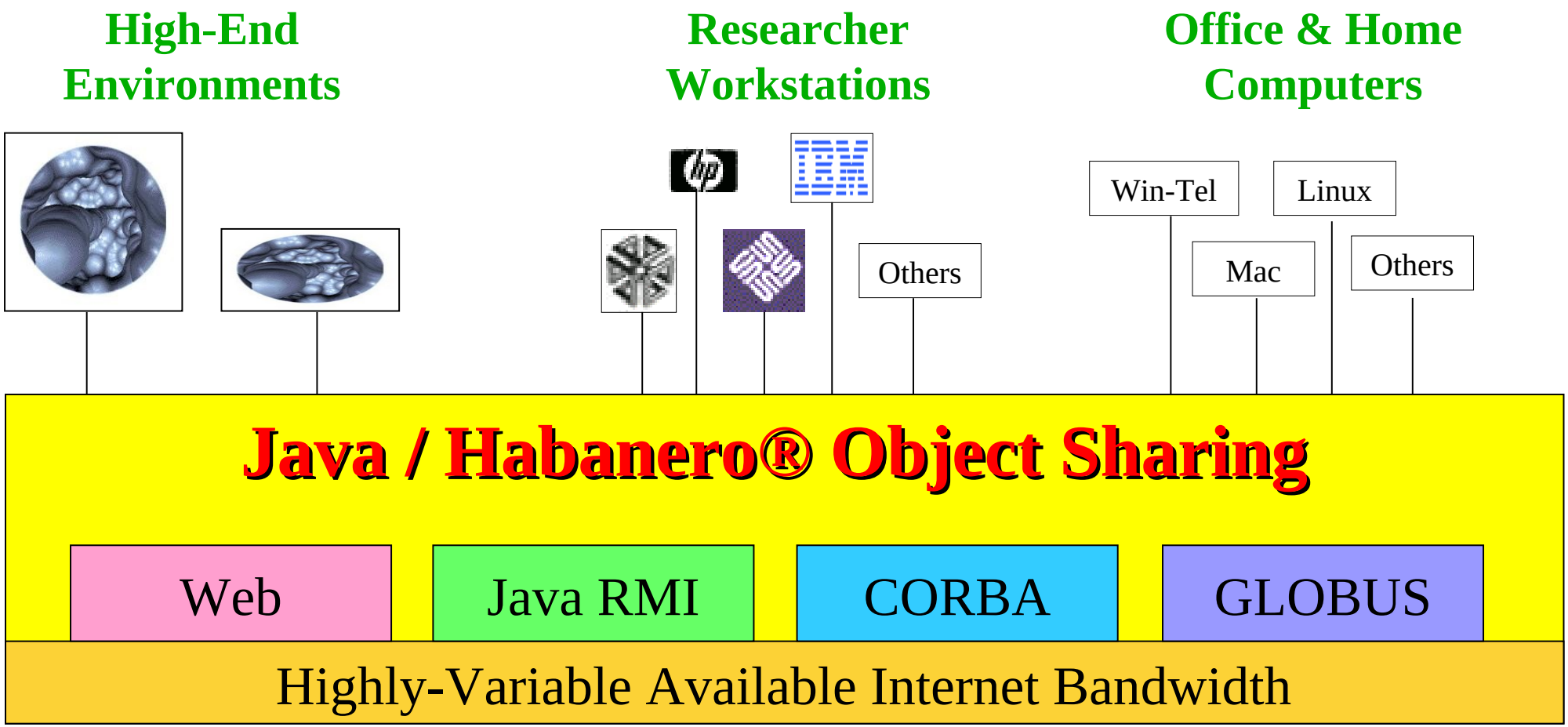
NASA IPG is Adding Funding To Collaborative Java3D

Coupling Data Formats to Visualization - NCSA's Hierarchical Data Format

- **HDF & Project Horizon**
  - Internet Access to Earth and Space Science Data
  - Science Data Browser (SDB)
    - To Provide Data Service for HDF & Other Formats
  - Java-based Viewers
    - Java-based HDF Browser
      - Standalone and Collaborative (Habanero™) Versions
    - General-purpose Image Viewer
- **HDF & ASCI**
  - The Data Models and Formats (DMF) Group
    - HDF As the Open Standard Exchange Format and I/O Library
  - ASCI HDF Requirements
    - Must Support Large (> a Terabyte) Datasets
    - Must Handle ASCI Data Types, Especially Meshes
    - Must Perform Well in Massive Parallel Environments
    - Store Unstructured Data for Efficient Visualization

Vision of the Java/Collaborative Future

- "Everybody Benefits" From HPC Science



Source: Larry Jackson, NCSA

Alliance Distance Education - Using JAVA Plug-ins to Web Browsers

- Universities "specialize" and deliver courses in areas of expertise
  - Provide all students *and faculty* with access to broader range of leading-edge courses
- Use Tango/WebWisdom for synchronous delivery of lectures from SU to JSU
- Pilot for distance *training* of DoD users.



Source: Geoffrey Fox, NPAC/Syracuse; DoD Army CEWES

Goal-Create Collaborative Interface to Link Multiple Investigators With the Grid

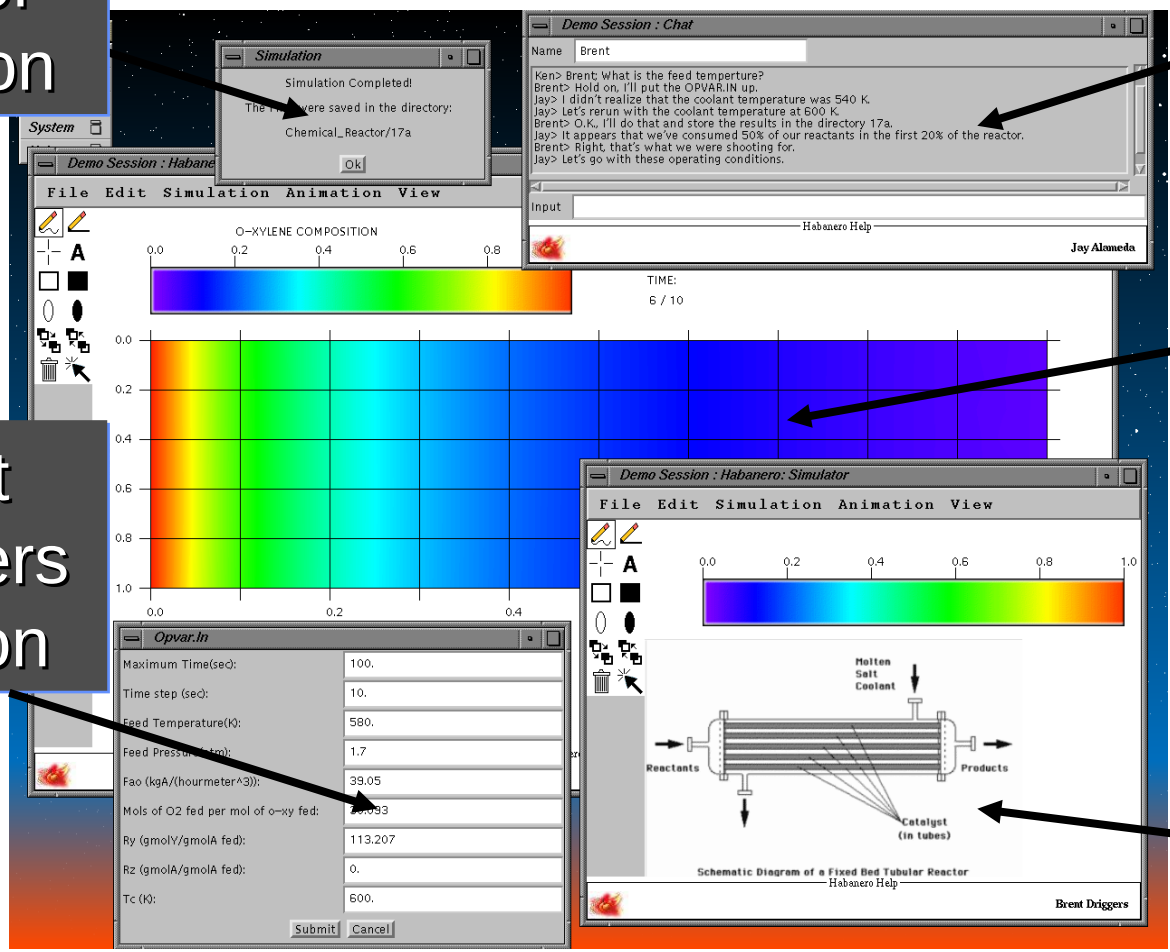
Status of Simulation

Interactive Discussion

Detailed Visualization

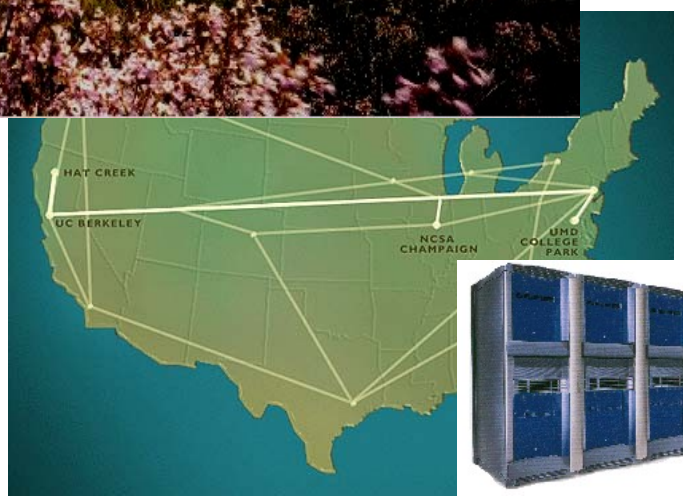
Current parameters in solution

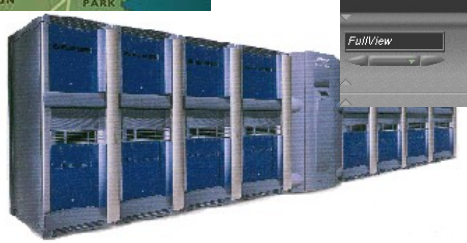
Reactor Simulation

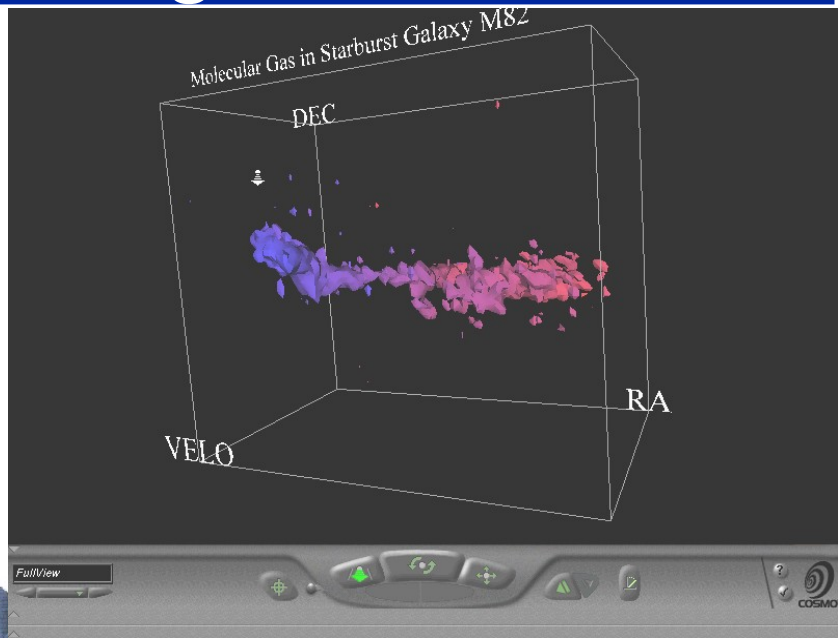


The Grid Links Remote Sensors With Supercomputers, Controls, & Digital Archives





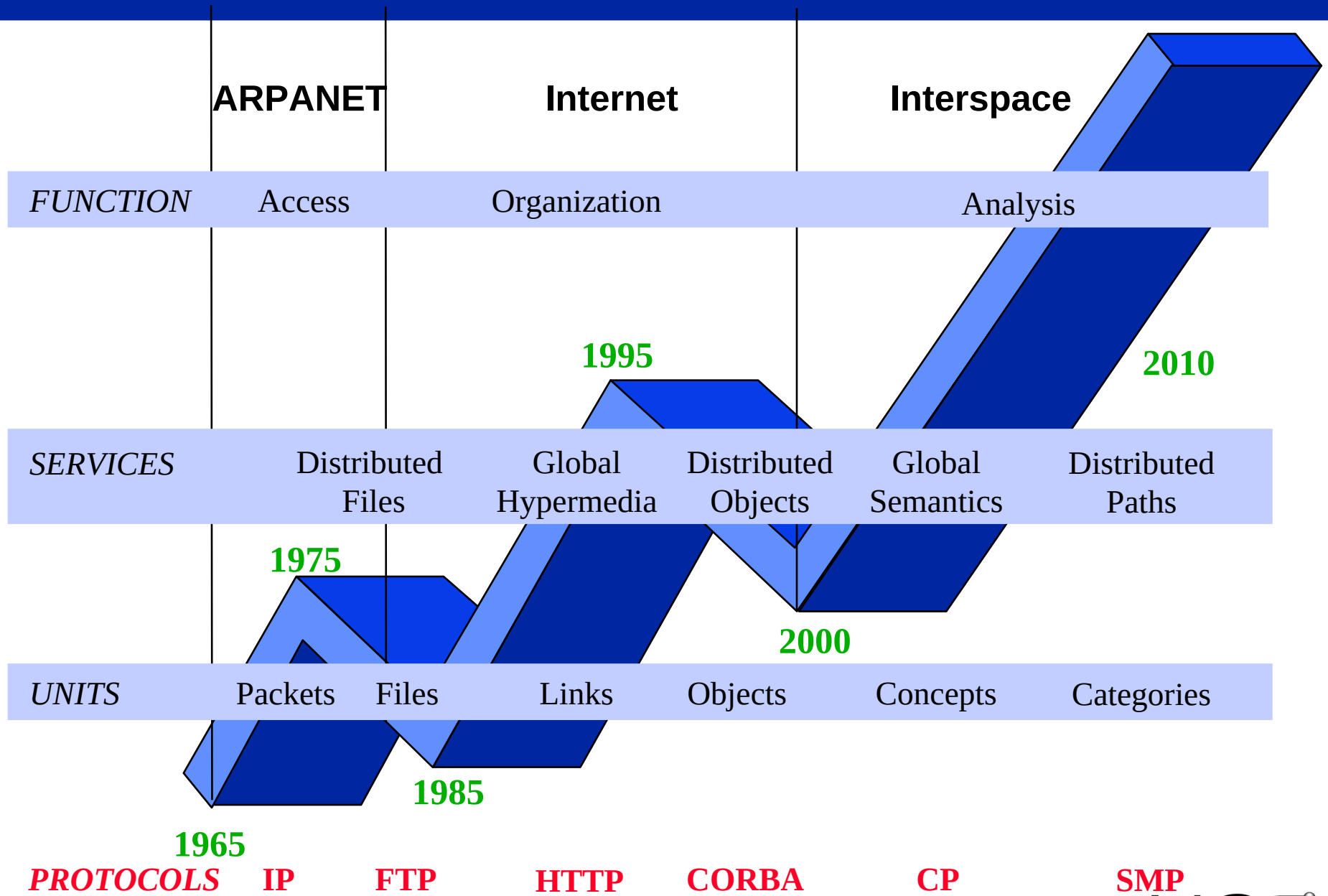




Starburst Galaxy M82

- **Alliance Scientific Instrument Team**
  - **Radio Astronomy and Biomedicine**
  - **Collaborative Web Interface**
  - **Real Time Control and Steering**

The Third Wave of Net Evolution

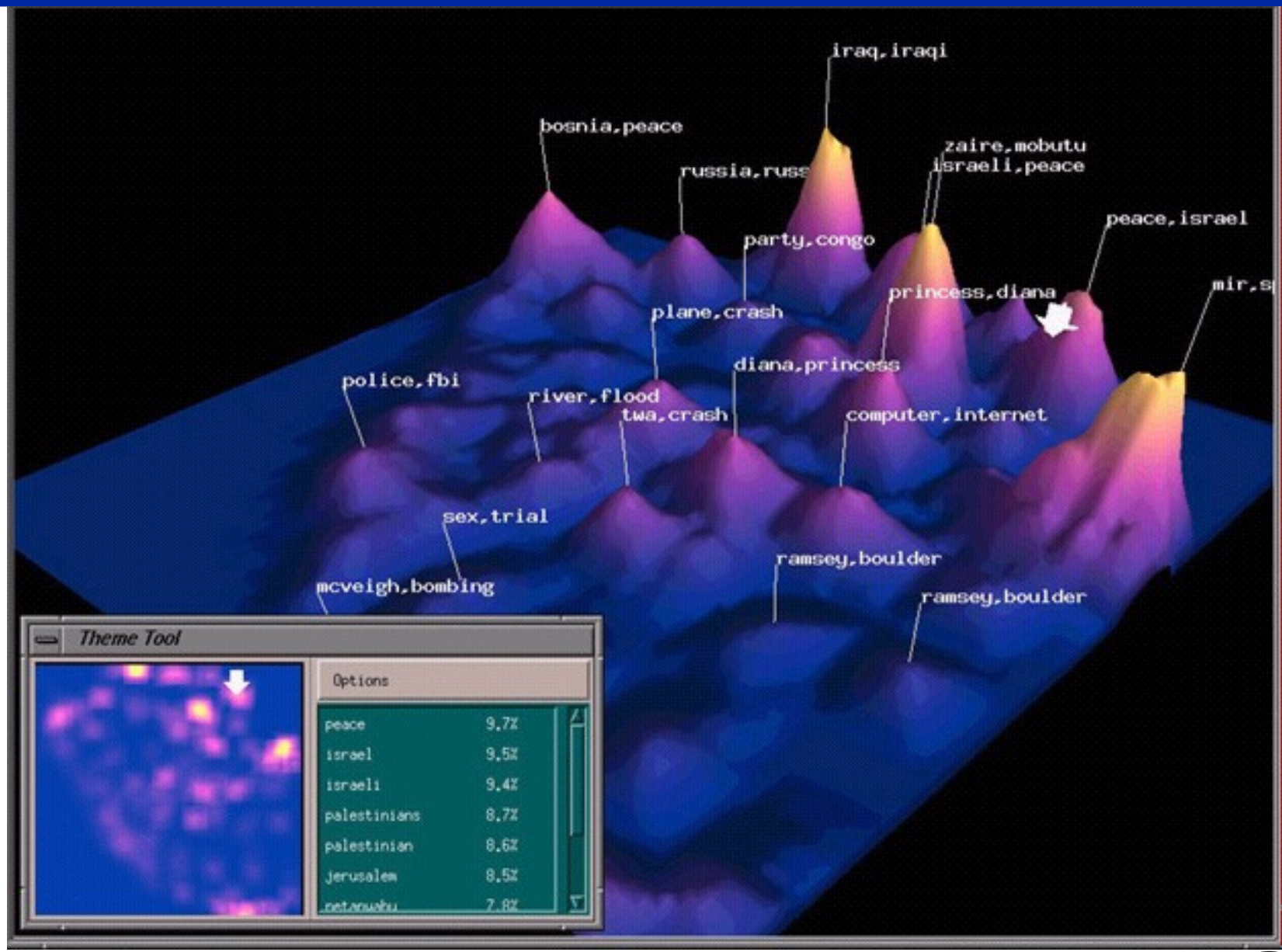


NCSA / UIUC Digital Library Initiative: Towards Scalable Semantic Retrieval

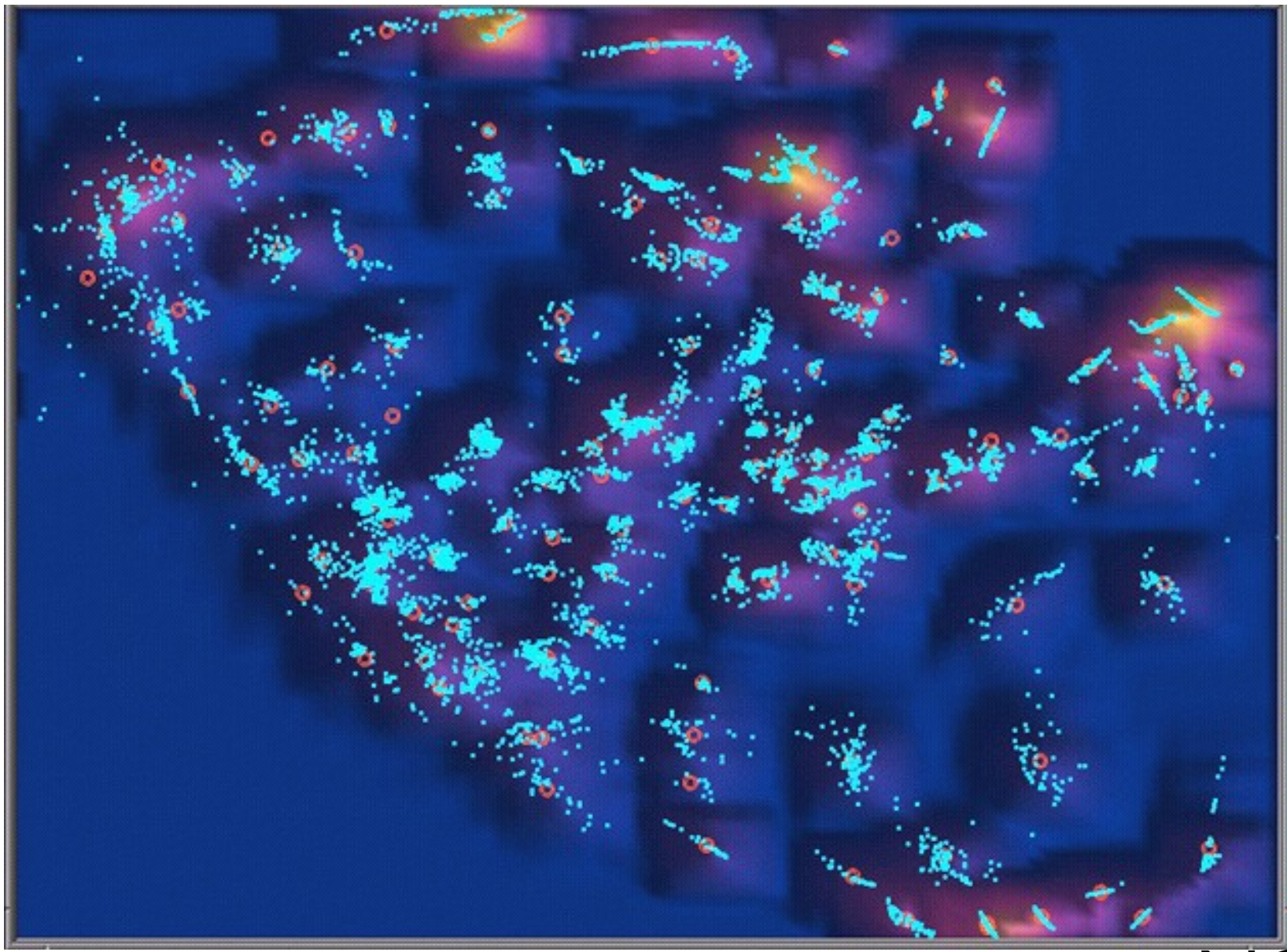
- Bruce Schatz, UIUC and Hsinchun Chen, U Arizona
- Automatic Indexing of Concepts
  - Find Context of Phrases within Documents
  - Concept Space Based on Term Frequency
- Useful for Interactive Searching
  - Given a Term, Can Suggest Other Terms
  - Concept Spaces Supports Vocabulary Switching
- Concept Spaces Require Supercomputing
  - Inspec Space (400K abstracts)
    - 1 day on 16-node SGI Challenge
  - 575 Spaces for Compendex (4M abstracts)
    - 3 days on 48-node HP Convex Exemplar

Science: June 7, 1996 and January 17, 1997

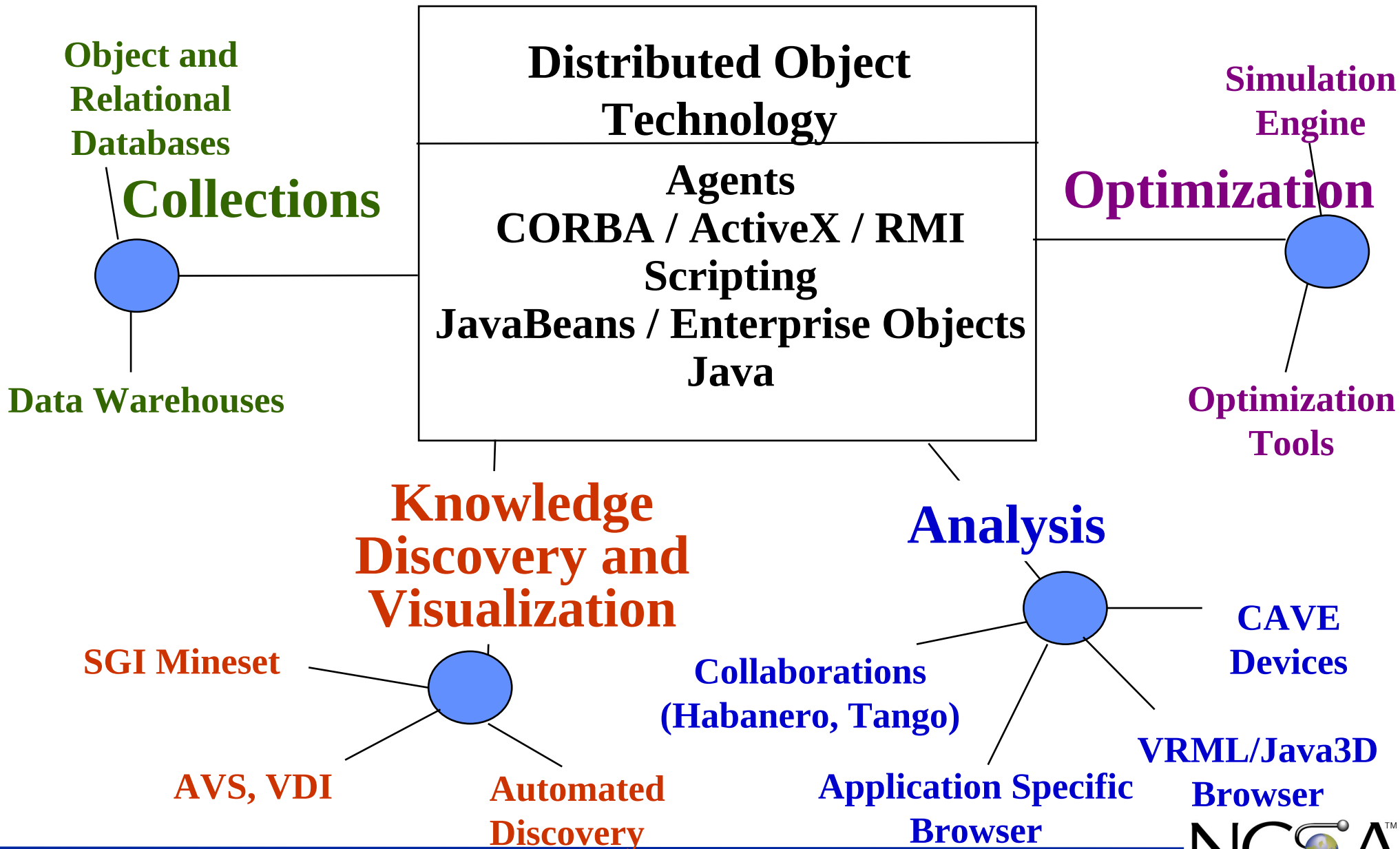
Visualizing Relationships Between Documents-
6500 News Stories from the WWW in 1997



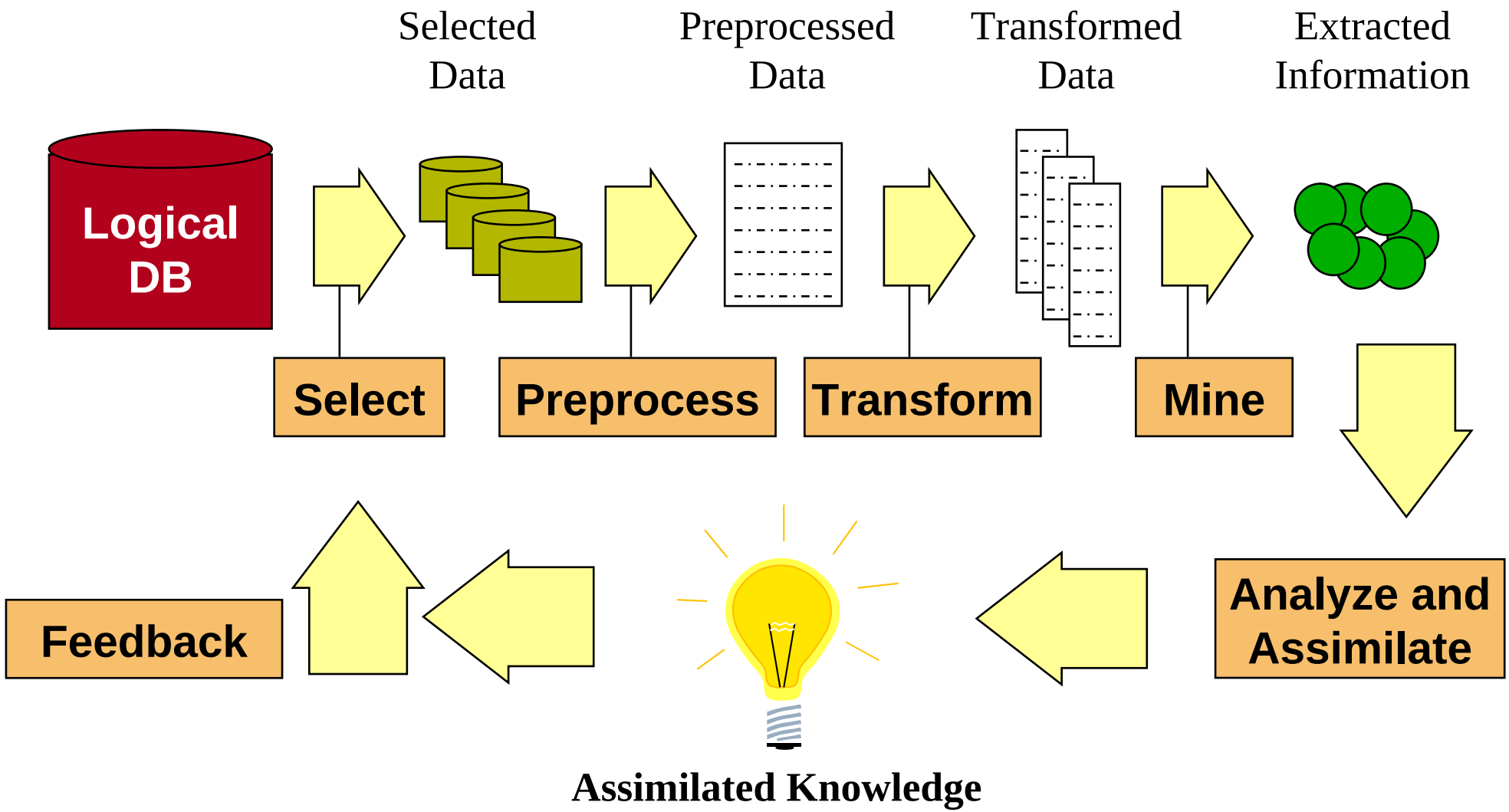
Visualizing Relationships Between Documents - Need Extension to Millions of Web Documents



NCSA Knowledge Management Workspaces



Knowledge Discovery Process



Automated Discovery and Learning - NCSA Techniques

- **Automated Discovery Tools**
  - Creation of Predication and Classification Models
  - Link Analysis
  - Deviation Detection
  - Database Segmentation
- **Automated Learning Research Topics**
  - Automatic Text Document Classification
  - Knowledge Source Integration
  - Parallel Algorithms for Induction
  - Interactive Self-organizing Maps

Automated Discovery By Machine Learning

- Creation of Prediction & Classification Models
  - Past Data Predicts Future Response
  - Typical Technique: Supervised Learning
    - Neural Nets
    - Decision Trees
    - Naïve Bayesian
- Link Analysis
  - Discover Relations Between Records in Datasets
    - Association
    - Sequential Pattern
    - Similar Time Sequence
  - Typical Techniques: Genetic Algorithms

Automated Discovery By Machine Learning

- **Database Segmentation**
  - Regroup Information Sets
    - Neural Clustering
    - Similar Characteristics, eg.Demographic Clustering
  - Typical Technique: Unsupervised Learning
    - SOM (Self-organizing Maps)
    - K-Means
- **Deviation Detection**
  - Identify Outliers in a Data Sample
    - Visualization
  - Typical Techniques: Stochastic Model Analysis
    - Probability Distribution Contrasts
    - Statistical Model Determination

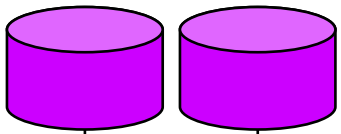
Data Mining - NCSA Industrial Partner Projects

- **Caterpillar**
  - Effluent Quality Control
  - Smart Selling
  - Warranty Claims Analysis
  - Customer Value Analysis
- **Ford**
  - Product Compatibility
  - Harshness, Noise, Vibration
  - Marketing
- **Sears**
  - Transaction Management
- **Boeing**
  - Post-Flight Diagnostics
- **Allstate**
  - Medical Claims

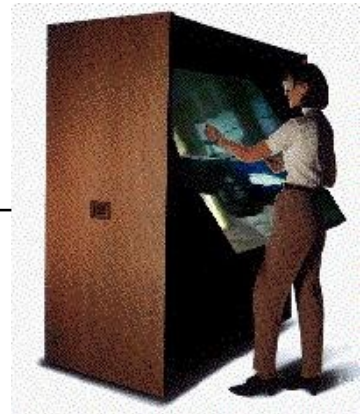
- **Financial Impact May Be Greater Than $30 Million**

NCSA Information Visualization Laboratory

Databases



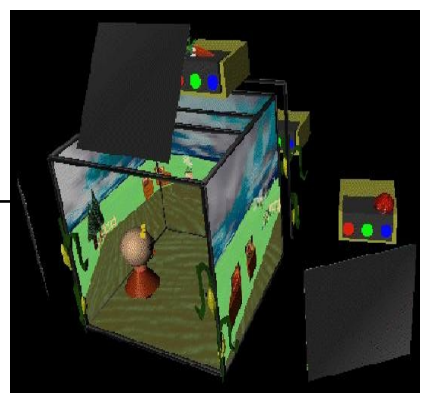
In3D™ for C++ and Java
VizIt/In3D™



Immersa Desk™

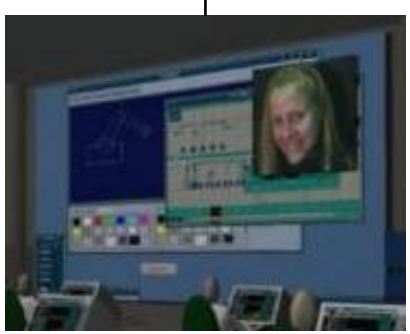


Graphics Workstations



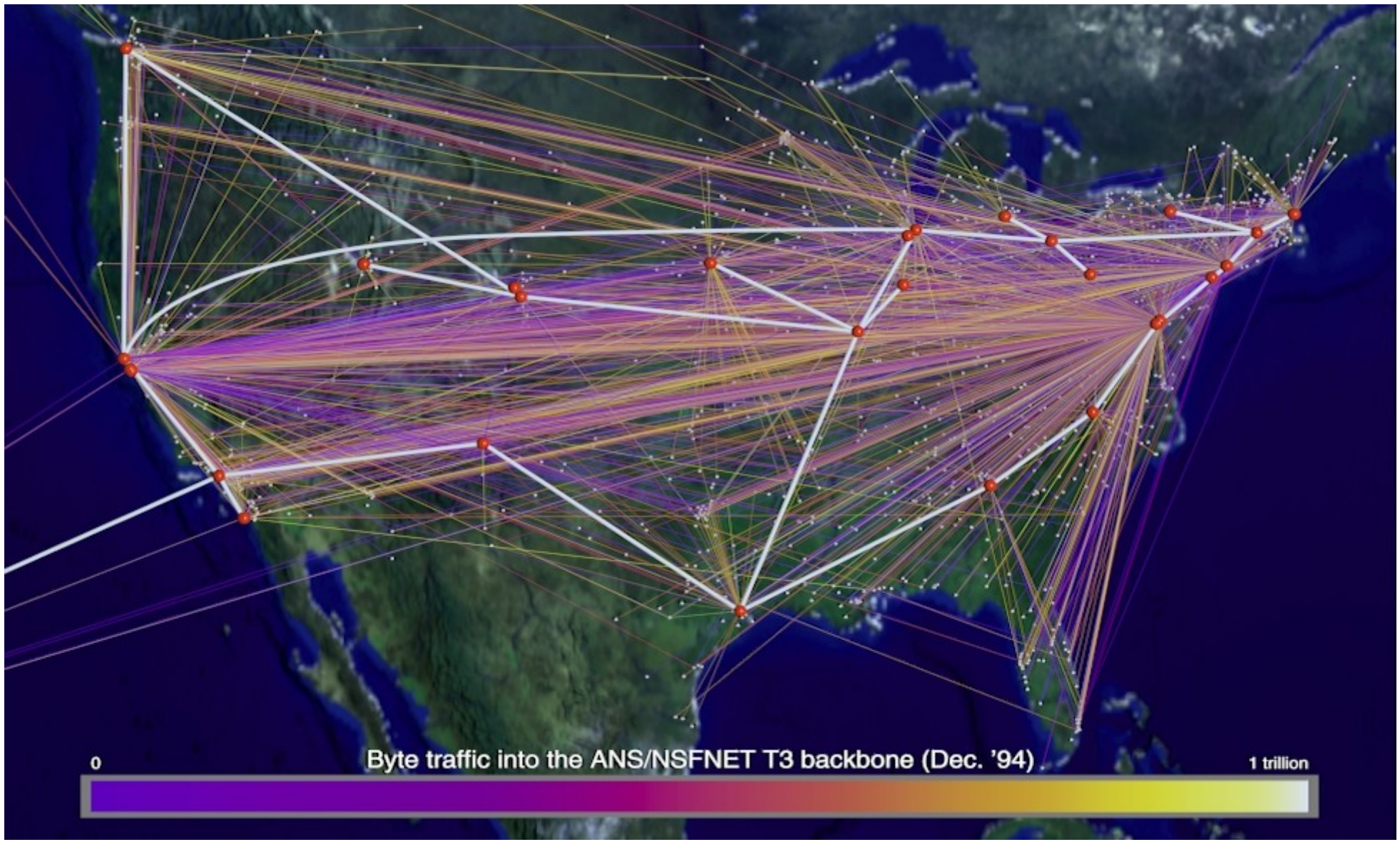
Cave™

MineSet
S-PLUS



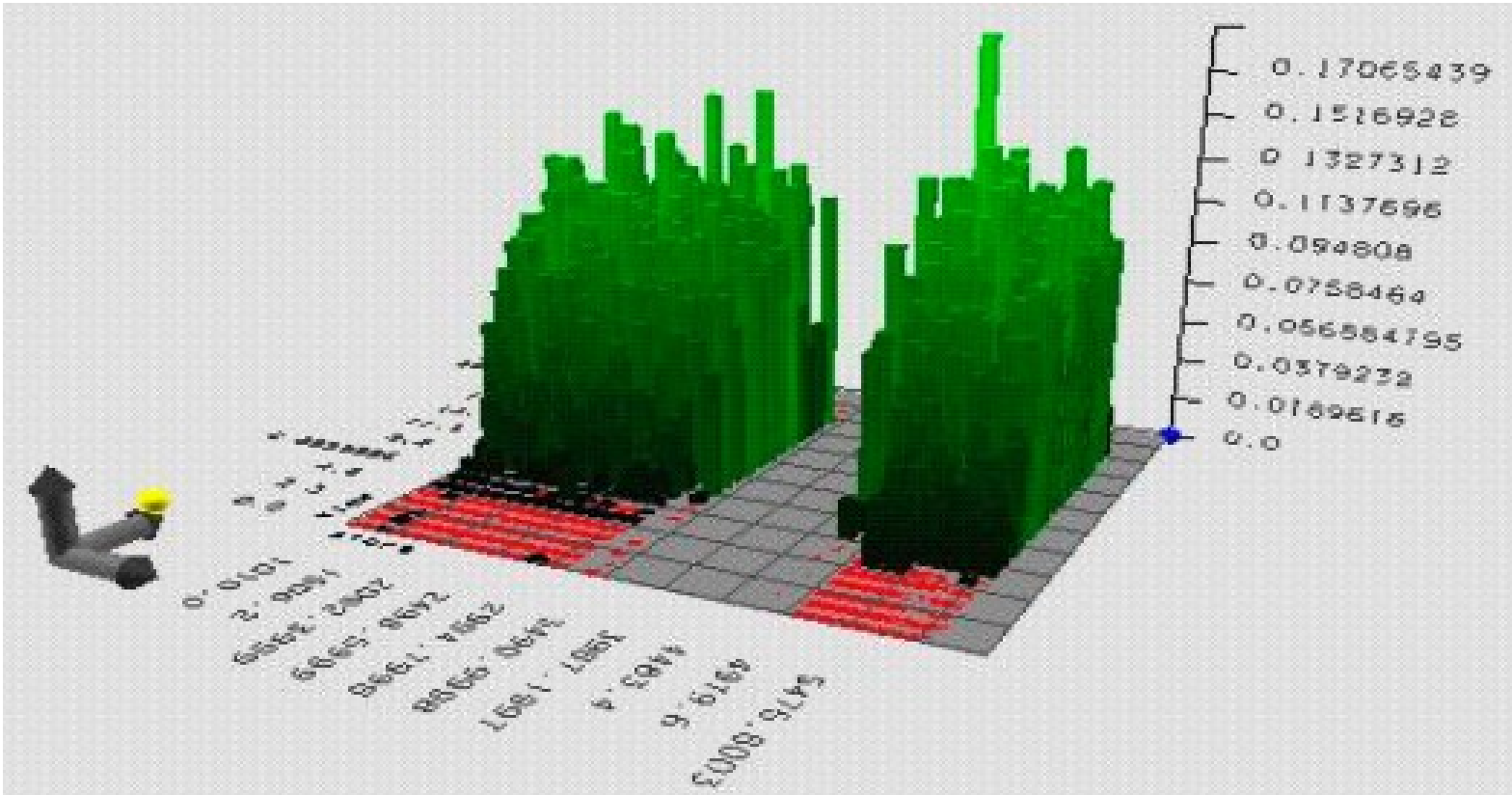
Flat Panel Wall

Information Visualization - Network Traffic



Robert Patterson, Donna Cox, NCSA

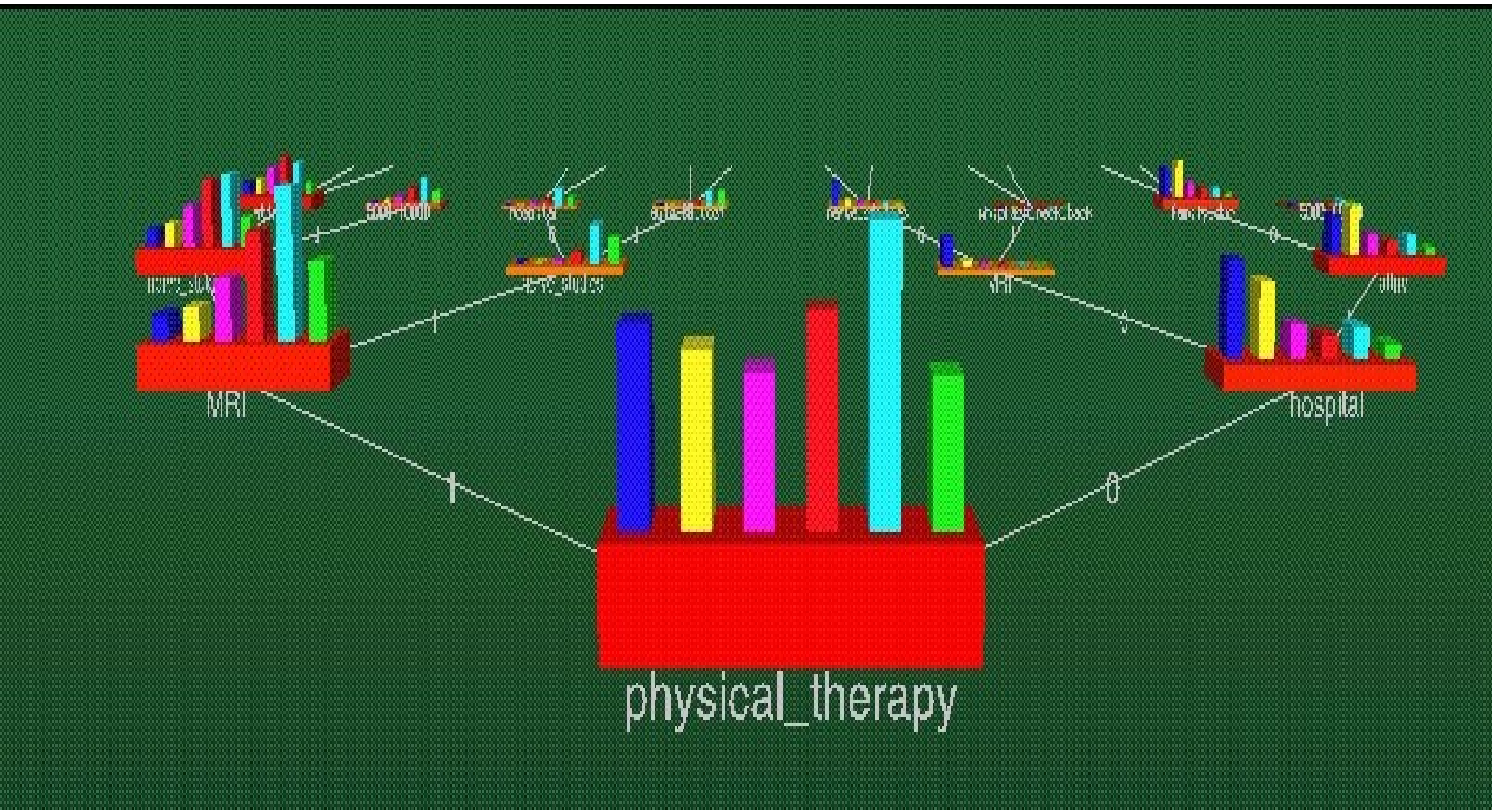
Sears Pioneers Massive Data Mining and Information Visualization at NCSA



- **1998 VLDB Survey Program Grand Prize Winner**
  - **Largest Database**
    - **4.7 Terabytes of Data**
  - **10 Terabyte Total Disk Space Capacity**
    - **Storage Provided by EMC**

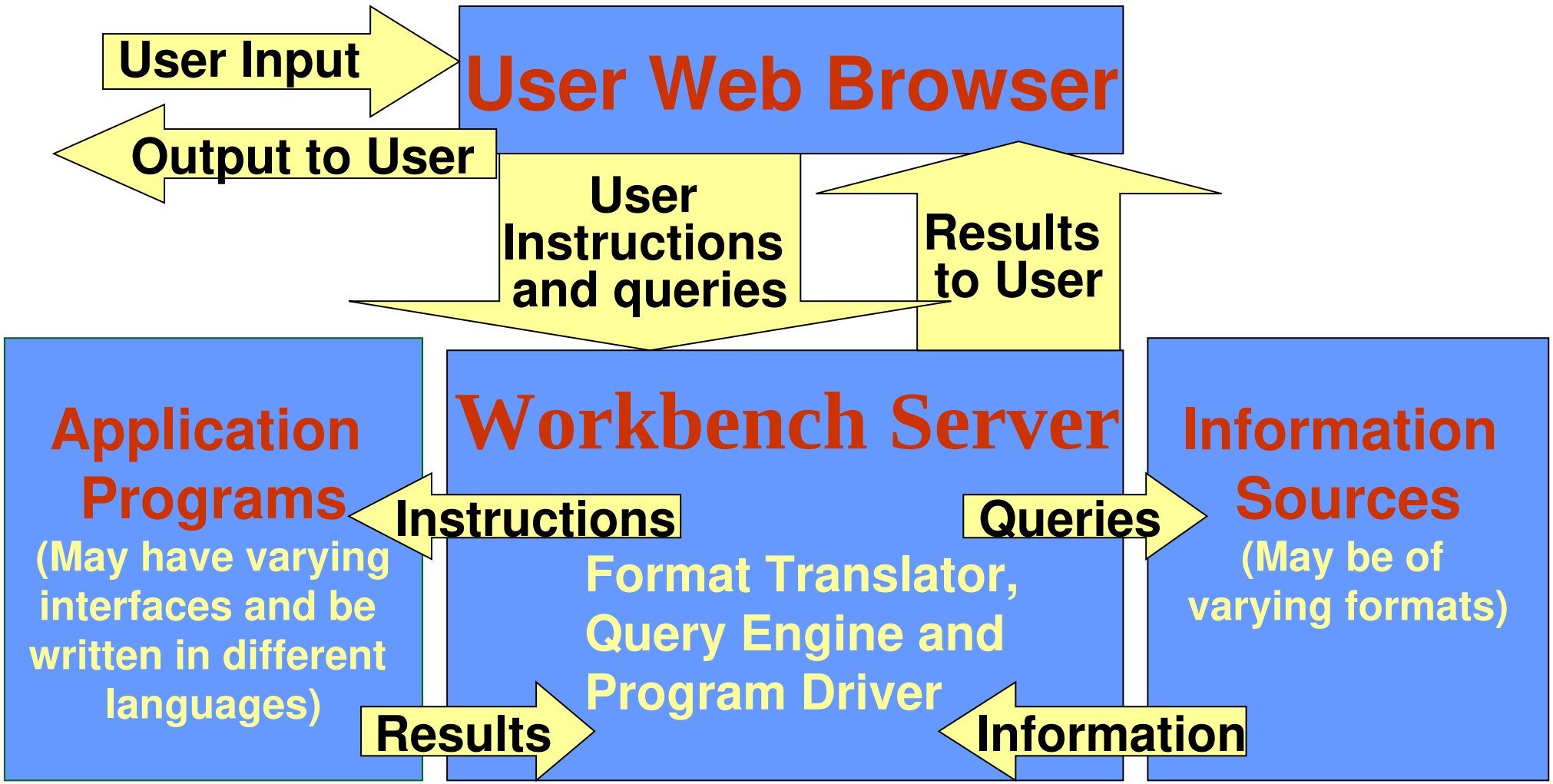
Image Courtesy of Michael Welge, NCSA and Sears

Information Visualization - Insurance Process Cost Drivers

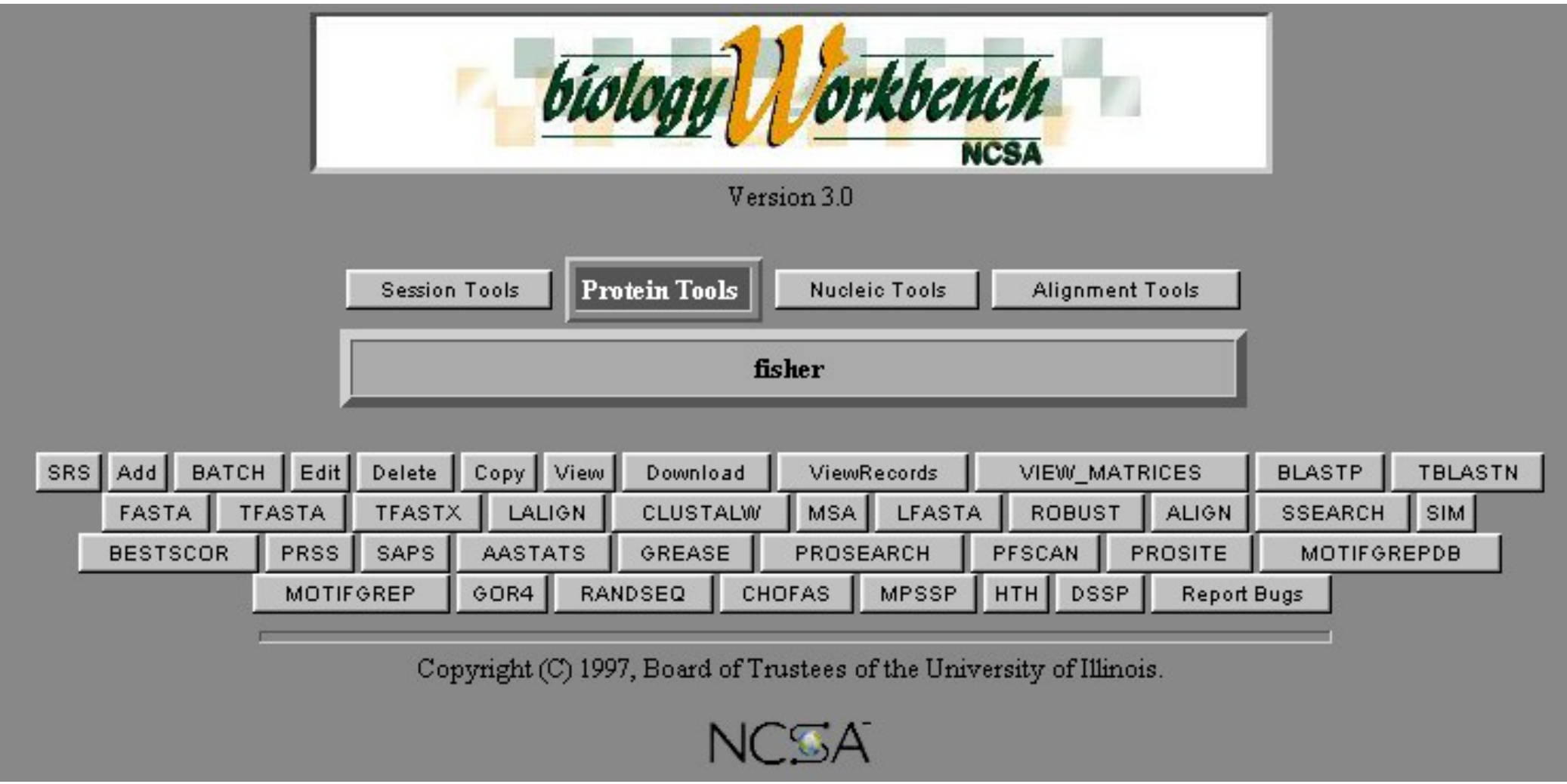


Automated Discovery Using SGI MineSet

The NCSA Information Workbench - An Architecture for Web-Based Computing



The NCSA Biology Workbench - Web Computing with Distributed Datasets



Powered by SGI Origin Supercomputer

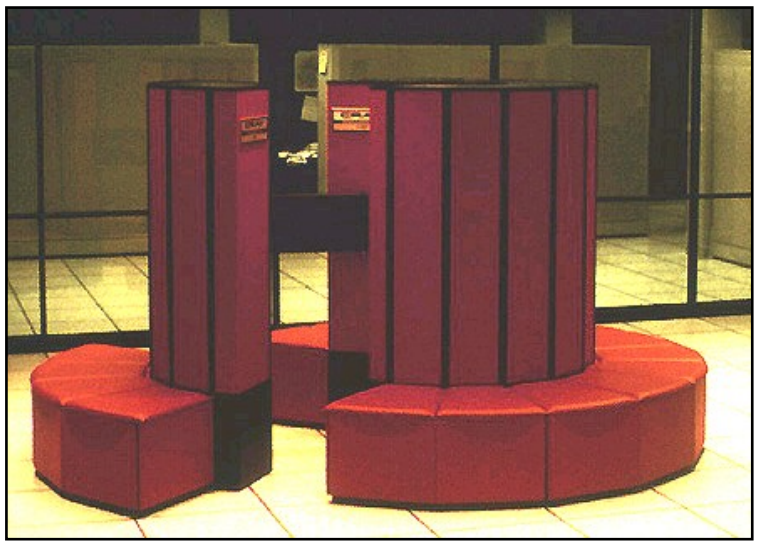
http://biology.ncsa.uiuc.edu/

NCSA™
National Computational Science Alliance

Toward a Social Sciences Workbench

- Potential New Project with Alliance
- Partner with ICPSR?
- Web Interface to Social Science:
  - Programs
  - Data

The Continuing Exponential Agent of Change



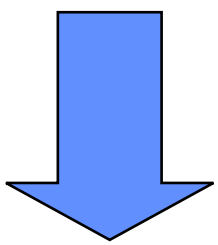
1985 Cray X-MP

Cost: $8,000,000

60,000 watts of power

No Built in Graphics

56 kbps NSFnet Backbone





1997 Nintendo 64

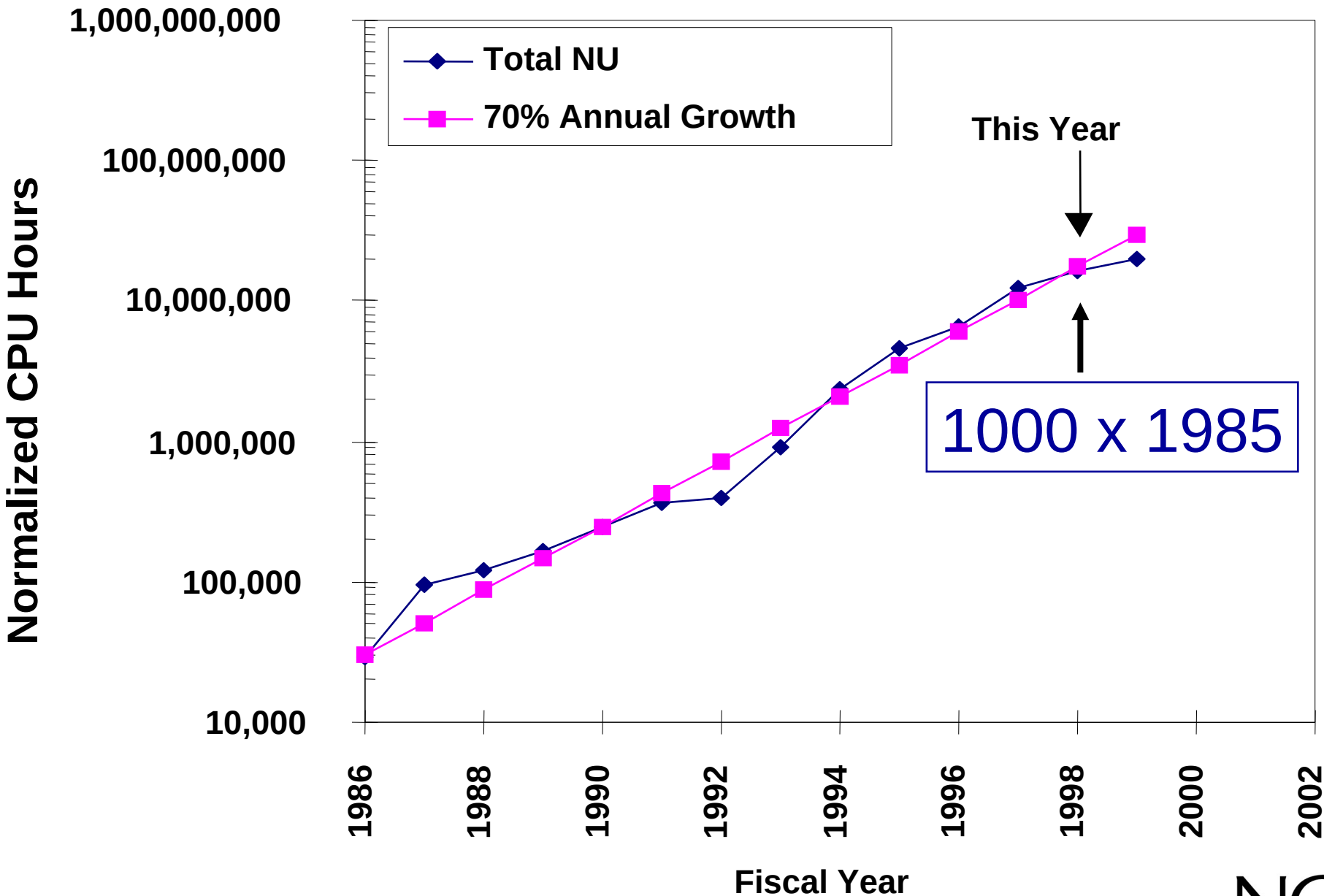
Cost: $149

5 watts of power

Interactive 3D Graphics

64 kbps ISDN to Home

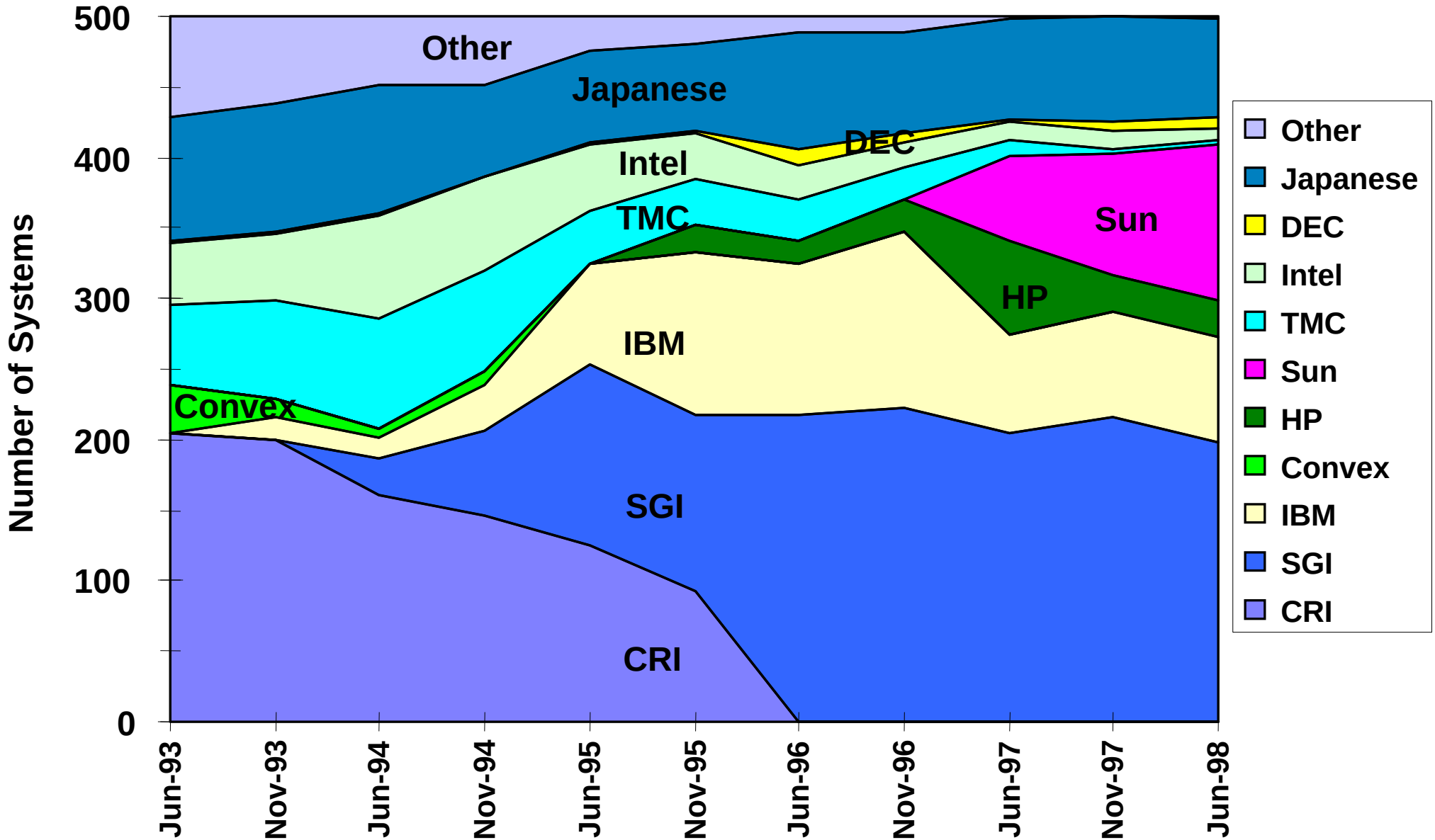
Growth Rate of the NSF Supercomputer Capacity is 70% Compounded Per Year!



Source: Quantum Research; Lex Lane, NCSA

National Computational Science Alliance

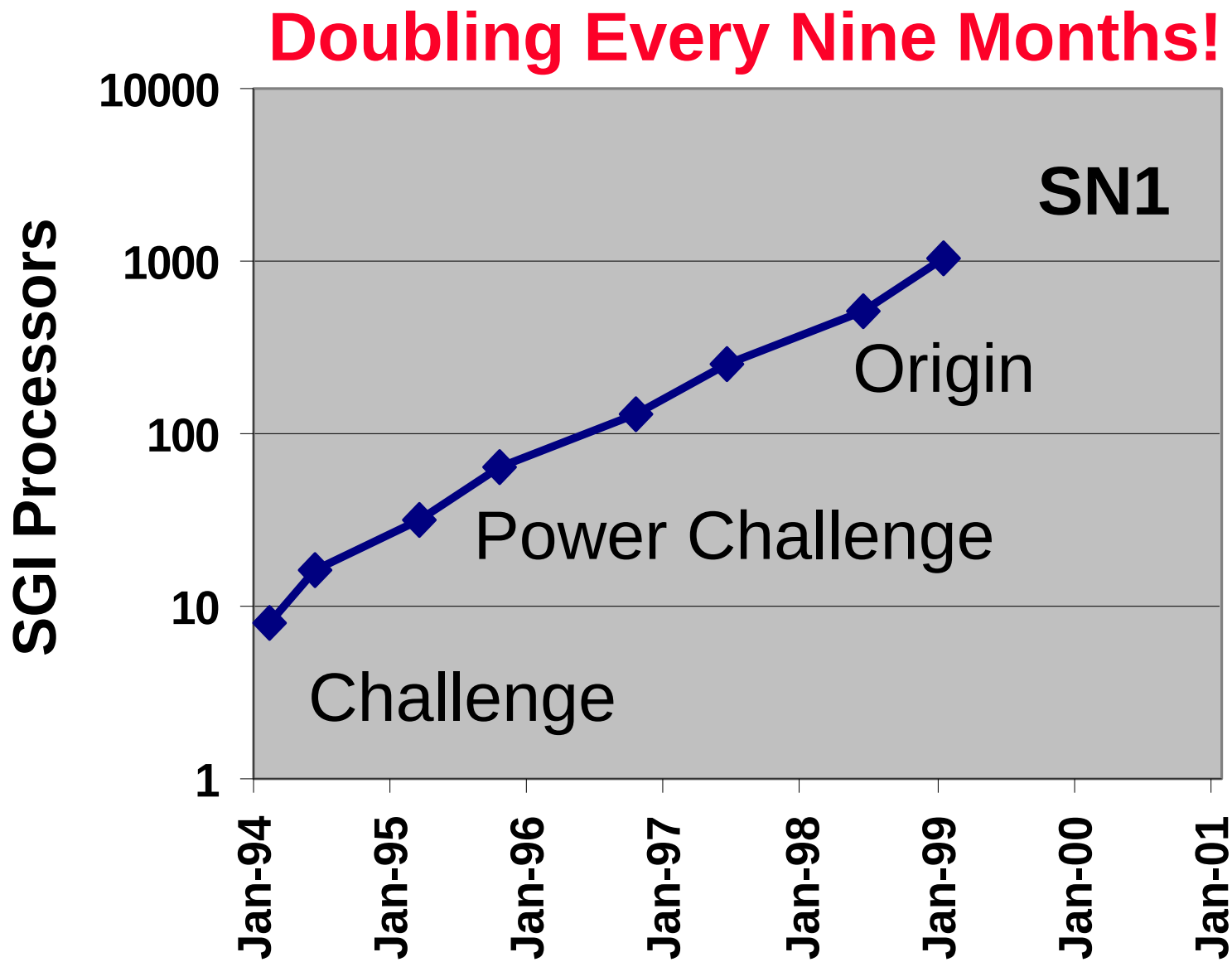
TOP500 Systems by Vendor - A Market Revolution



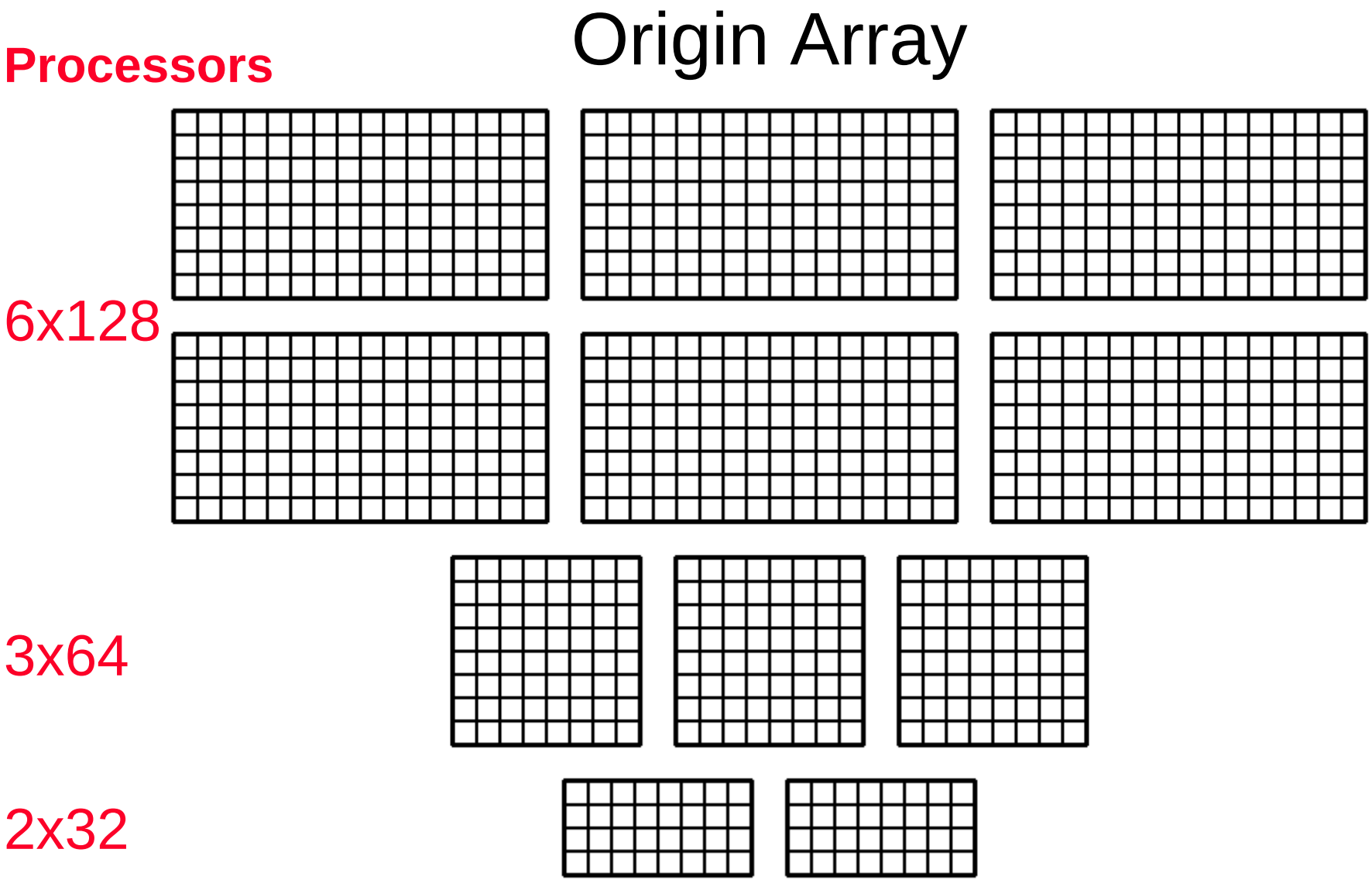
TOP500 Reports: http://www.netlib.org/benchmark/top500.html

National Computational Science Alliance

NCSA is Combining Shared Memory Programming with Massive Parallelism



Proposed NCSA Silicon Graphics Cray Origin Array - 1024 Processors



Subject to NSF Approval of Funds

National Computational Science Alliance

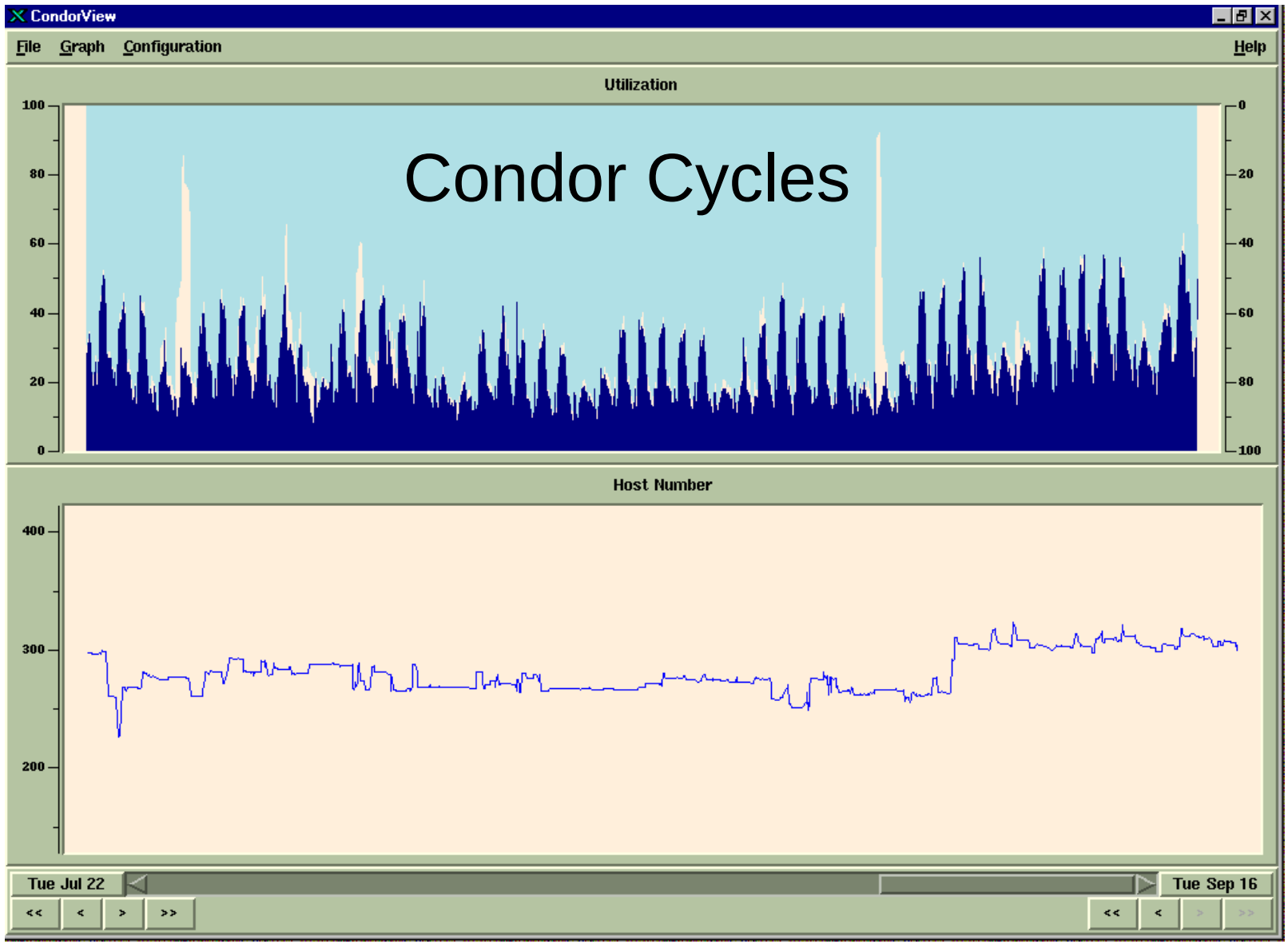
JP Morgan Hero Calculation

- HPC Strategic Business Analysis
- Calculations Used 128-Processor SGI Origin
  - Two Week Period in January 1998
  - NCSA and SGI Doubled Memory in a Week
- Extended JPM's Risk Management Capabilities
- Hundreds of Market Scenarios Simulated
- NCSA, Strategic Vendor, Industrial Partner
  - Existing Relationships Facilitated Quick Startup
  - Win-Win-Win Result

Challenge-How to Increase the Number of Social Scientists Using High Performance Computing?

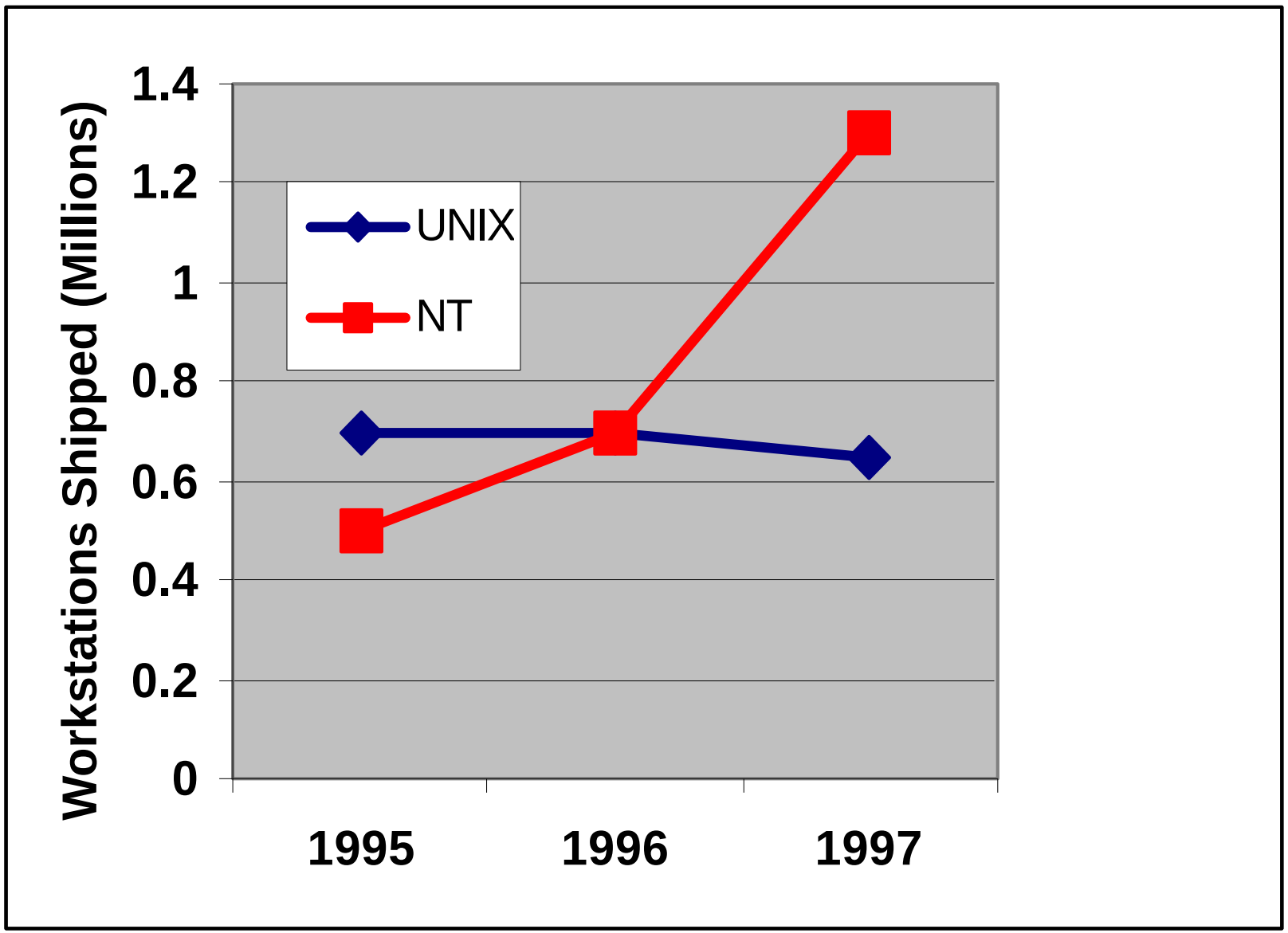
- **NSF Supercomputer Centers in FY97**
  - Consider All 900 Projects Using More Than 10 CPU-Hours
  - 7 out of 900 Projects Were Social Science
- **Social Science Project Areas**
  - Testing Time Series
  - Dynamic Optimization
  - Large Scale GIS
  - Economics
    - Competitiveness Models and Strategies
    - Economic Behaviour
    - Capital Structures
    - Stock Market Models

Computing on the University of Wisconsin Condor Pool



CondorView, Courtesy of Miron Livny, Todd Tannenbaum(UWisc)

NT Workstation Shipments Rapidly Surpassing UNIX



Source: IDC, Wall Street Journal, 3/6/98

The University of Illinois NT Supercluster - 256 Intel Pentium II Processors

“Supercomputer performance at mail-order prices”-- Jim Gray, Microsoft

- **Andrew Chien, Computer Science UIUC**
- **Rob Pennington, NCSA**

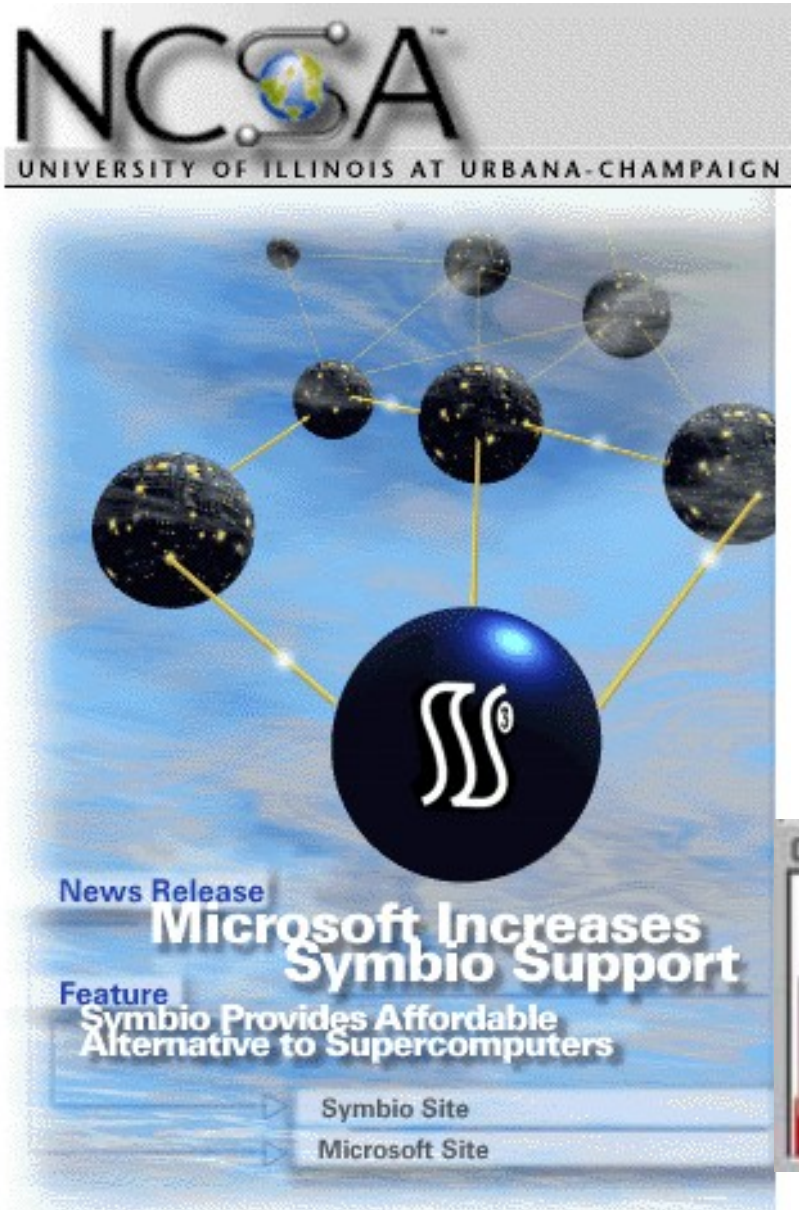


192 Hewlett Packard
300 MHz

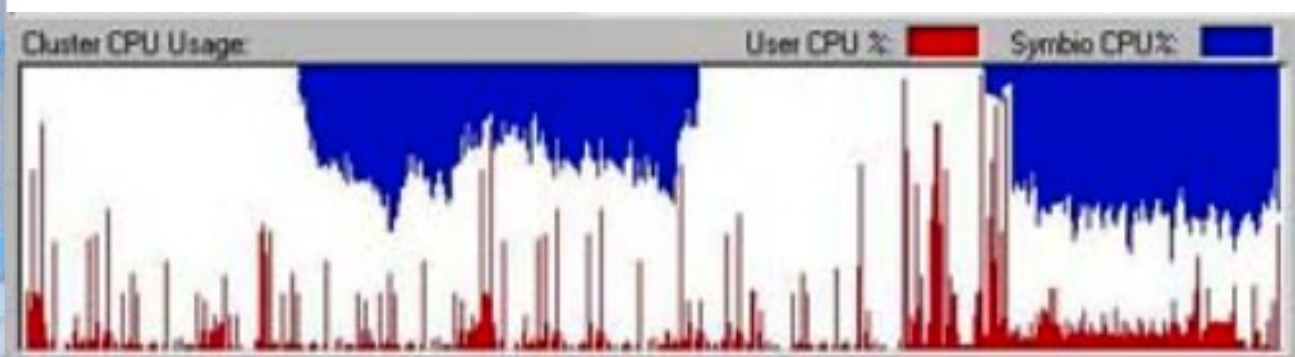


64 Compaq 333 MHz

NCSA Symbio - A Distributed Object Framework Bringing Scalable Computing to NT Desktops



- **Parallel Computing on NT Clusters**
  - Briand Sanderson, NCSA, Microsoft
  - Microsoft Co-Funds Development
- **Features**
  - Based on Microsoft DCOM
  - Batch or Interactive Modes
  - Application Development Wizards
- **Current Status & Future Plans**
  - Symbio Developer Preview 2 Released
  - Princeton University Testbed



http://access.ncsa.uiuc.edu/Features/Symbio/Symbio.html

NCSA™
National Computational Science Alliance

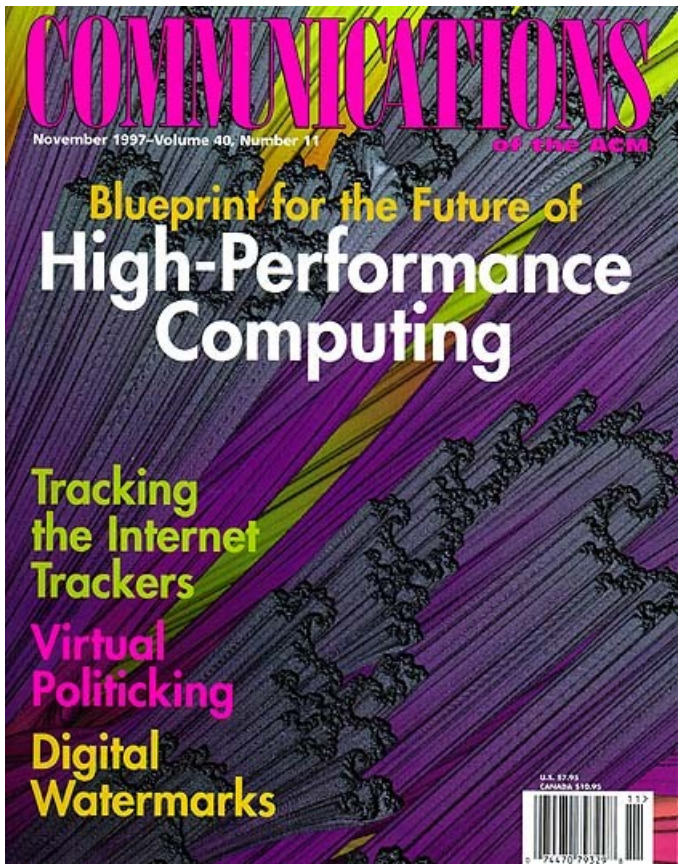
NSF / NCSA Federal Consortium

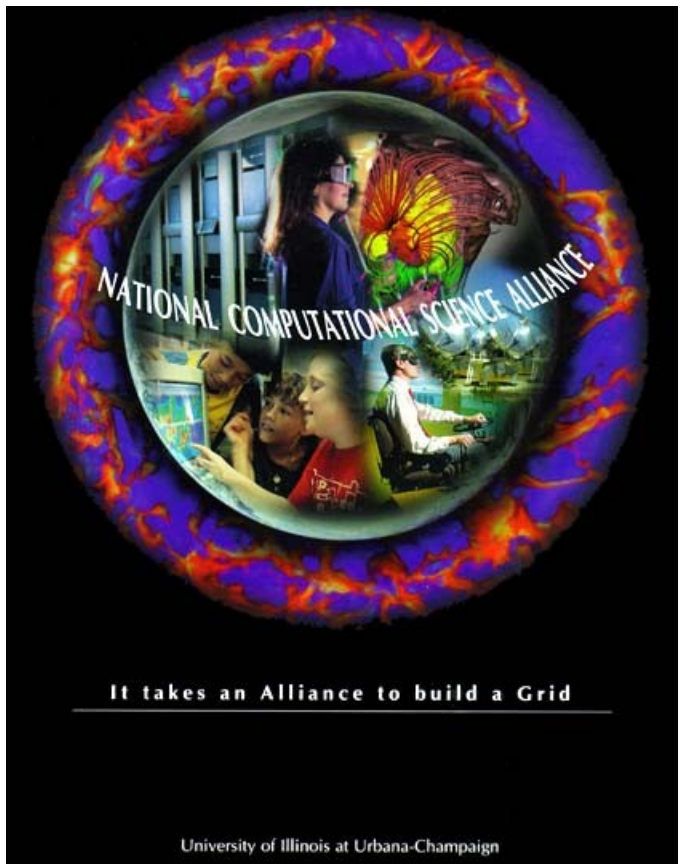
- **Member Agencies:**
  - Bureau of Census
  - Central Intelligence Agency
  - Defense Technical Information Center
  - Rural Development, Department of Agriculture
  - Department of Education
  - Department of Housing and Urban Development
  - National Biological Service
  - National Institutes of Health
  - National Oceanic and Atmospheric Administration
  - NASA
  - National Science Foundation
  - National Security Agency
  - Nuclear Regulatory Commission

- **Funding IT Development**
  - Security
  - Universal Access
  - Distance Learning
  - Intranet Technology
- **Staff Training**
- **Electronic Meeting Spaces**



How to Find Out More About the Alliance





See also http://alliance.ncsa.uiuc.edu