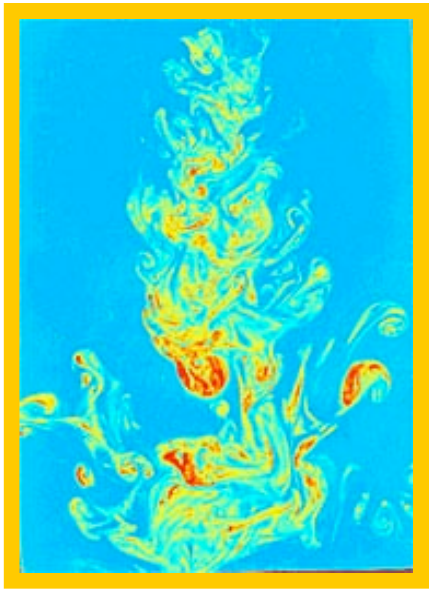


Cyberinfrastructure-enabled Research and Education at SDSC

Dr. Francine Berman

Director, San Diego Supercomputer Center

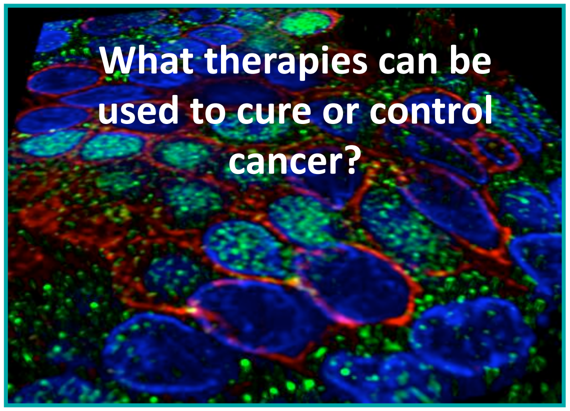
Professor and High Performance Computing Endowed Chair, Department of Computer Science and Engineering, UC San Diego

Why Cyberinfrastructure-enabled Research Matters:

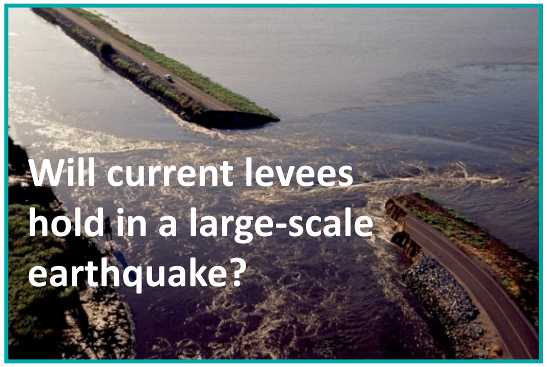
21st Century Problems Require 21st Century Solutions

What is the impact of Global Warming?

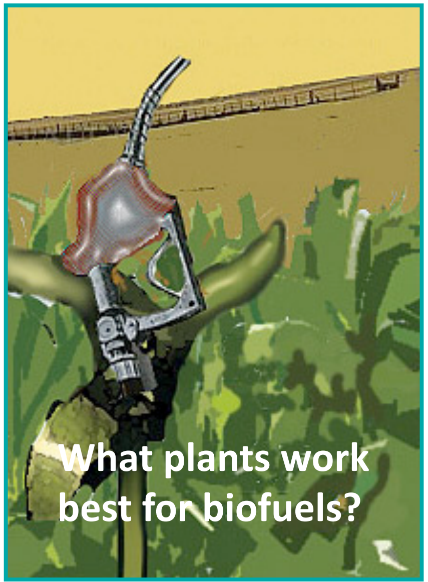
Can financial data be used to accurately predict market outcomes?



What therapies can be used to cure or control cancer?



Will current levees hold in a large-scale earthquake?



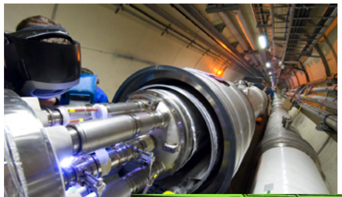
What plants work best for biofuels?

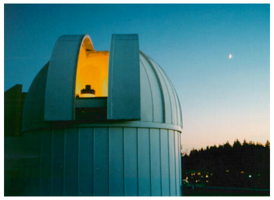
*"Let us be the generation that reshapes our economy to compete in the digital age. Let's set high standards for our schools and give them the resources they need to succeed. … **let's invest in scientific research,** and let's lay down broadband lines through the heart of inner cities and rural towns all across America."*

Barack Obama

Cyberinfrastructure is the foundation for modern research and education

- **Cyberinfrastructure components:**
  - Digital data
  - Computers
  - Wireless and wireline networks
  - Personal digital devices
  - Scientific instruments
  - Storage
  - Software
  - Sensors
  - People …





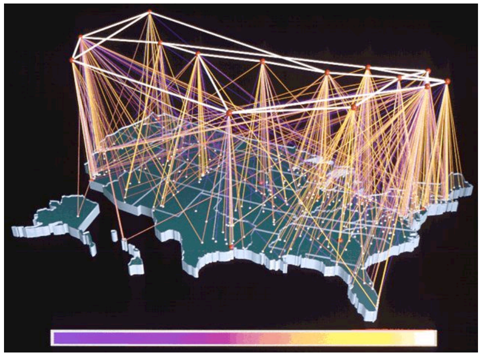
Cyberinfrastructure:
the organized aggregate of information technologies coordinated to address problems in science and society.

If infrastructure is required for an industrial economy, then we could say that cyberinfrastructure is required for a knowledge economy.”

NSF Final Report of the Blue Ribbon Advisory Panel on Cyberinfrastructure

SDSC's Challenge: *Accelerate Research and Education using Cutting-Edge Tools as the IT landscape changes*

Evolving IT landscape:



Late '80's - early 90's

Address Grand Challenge Problems using Supercomputers

Emerging technology: supercomputers

Mid-90's – mid 00's

Integrate diverse information technologies to solve modern challenges

Emerging technology: Grids

Mid 00's – present

Link researchers to "unlimited resources" in a way that facilitates their use to create useful information and new knowledge

Emerging technology: Clouds

Today's SDSC

Research and Education

- In 2007, SDSC was home to over 110 research projects involving researchers at UCSD and throughout academia.
- SDSC hosts hosted over 100 separate community digital data sets and collections for sponsors such as NSF, NIH, and the Library of Congress.
- SDSC staff and collaborators published scholarly articles in a spectrum of journals including *Cell, Science, Nature, Journal of Seismology, Journal of the American Chemical Society, Journal of Medicinal Chemistry, Nano Letters, PLoS Computational Biology*, and many others
- SDSC provided training and oversight for practica for over 200 students in 2007

Resources

- SDSC has ~250 research, technology, and IT staff with expertise in data use, management, and preservation, high performance computing, software tools, domain science, and education
- SDSC's data center is one of the largest academic data centers in the world with 36+ PBs capacity
- SDSC is home to the San Diego Network Access Point (SDNAP)
- SDSC's computers are architected to support data-intensive applications

Today's Presentation

- Cyberinfrastructure-enabled Applications at SDSC
  - Personalized Medicine
  - Simulating the Universe 1 Billion Years after the Big Bang
- SDSC Cyberinfrastructure Resources and Services
  - Triton
  - Chronopolis
- Cyberinfrastructure for the Next Decade

Promoting Good Health

- **How can we use information technologies to promote good health?**
  - – Better monitoring
  - – Better diagnosis
  - – Better cures
  - – Better therapies
  - – Better health infrastructure

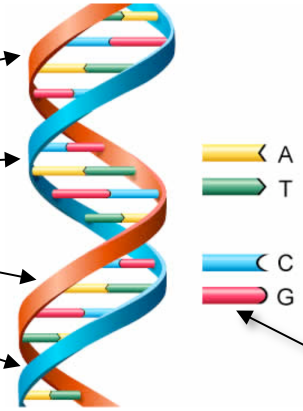
Personalized Medicine

- Current “$1,000” gene testing services examine only < 0.1% of the complete human genome
- SDSC’s Allan Snavely and his group working with medical researchers to develop high performance computational techniques to efficiently analyze the entire genome

What could you do if you knew your entire genome?

- Know if you are predisposed to certain diseases
- Customize medical treatments to be maximally effective for you

Current genotyping services only look for localized variations (“SNPs”) in a very small portion of the genome



True personalized medicine will require broad examination of the full genome of tens of thousands of people

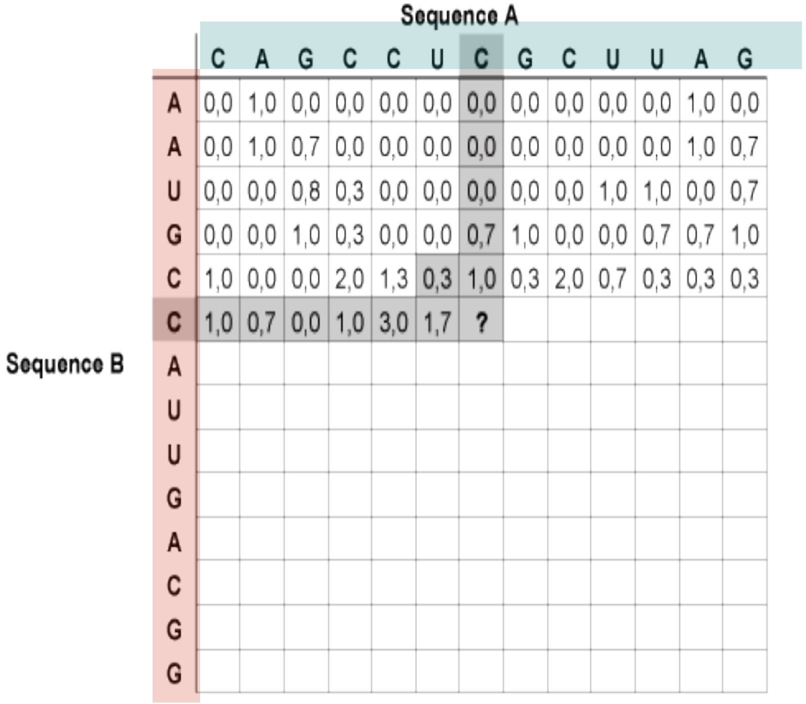
Human genome consists of 3,200,000,000 “base pairs” of data requiring massive computation for full analysis

Using Dynamic Programming to Compare Genetic Sequences

- Algorithms such as *Smith-Waterman*, used for genetic sequence comparison, are instances of dynamic programming

- Large search space structured into a succession of stages, such that
  1. the initial stage contains trivial solutions to sub-problems,
  2. each partial solution in a later stage can be calculated by recurring a fixed number of partial solutions in an earlier stage and
  3. the final stage contains the overall solution.

- Data stored in 3B+ X 3B+ base-pair matrix



Sequence A

Sequence B

	C	A	G	C	C	U	C	G	C	U	U	A	G
A	0,0	1,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	1,0	0,0
A	0,0	1,0	0,7	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	1,0	0,7
U	0,0	0,0	0,8	0,3	0,0	0,0	0,0	0,0	0,0	1,0	1,0	0,0	0,7
G	0,0	0,0	1,0	0,3	0,0	0,0	0,7	1,0	0,0	0,0	0,7	0,7	1,0
C	1,0	0,0	0,0	2,0	1,3	0,3	1,0	0,3	2,0	0,7	0,3	0,3	0,3
C	1,0	0,7	0,0	1,0	3,0	1,7	?						
A													
U													
U													
G													
A													
C													
G													
G													

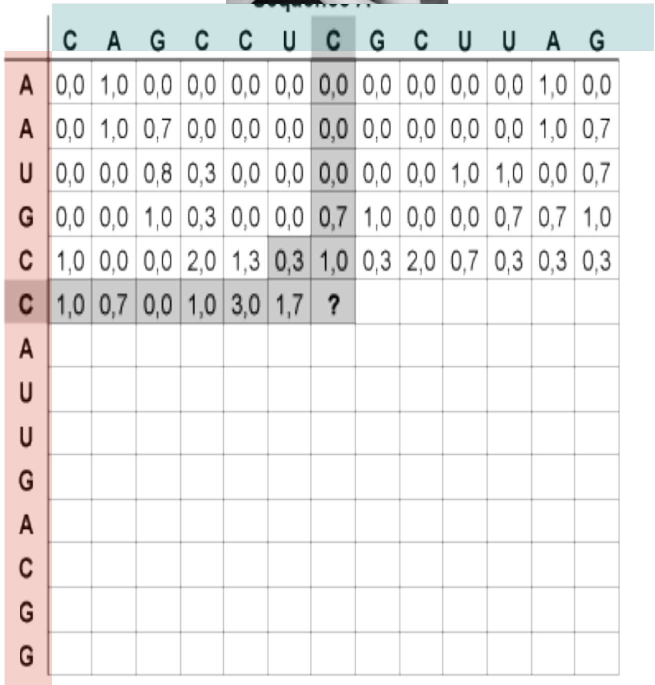
Genomic Analysis at Scale

Snavely's Group employing a multi-pronged approach

1. **Create new algorithms that can efficiently calculate structural similarities** (the "architectural difference") between two sequences in parallel on tightly-coupled HPC architectures

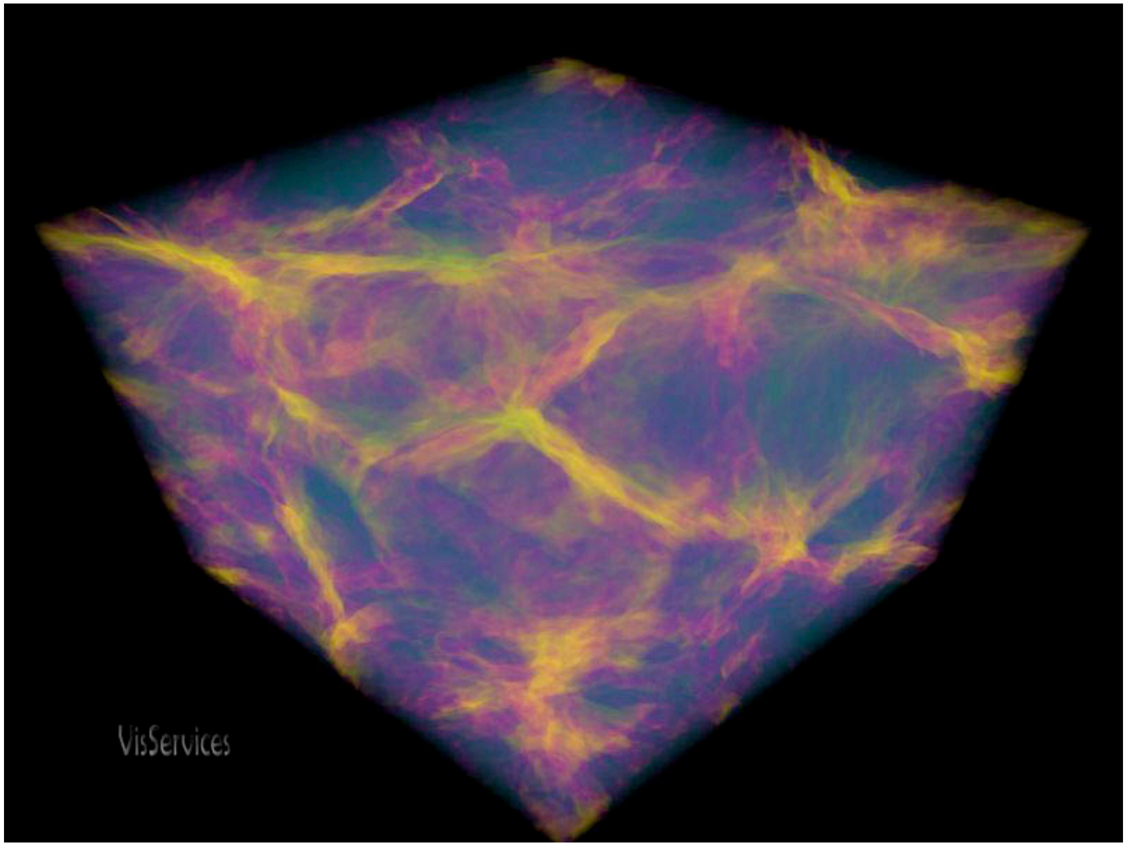
2. **Create new algorithms that tolerate high latency, high processor count cloud environments**

3. **Create new approaches for querying and storing very large databases**



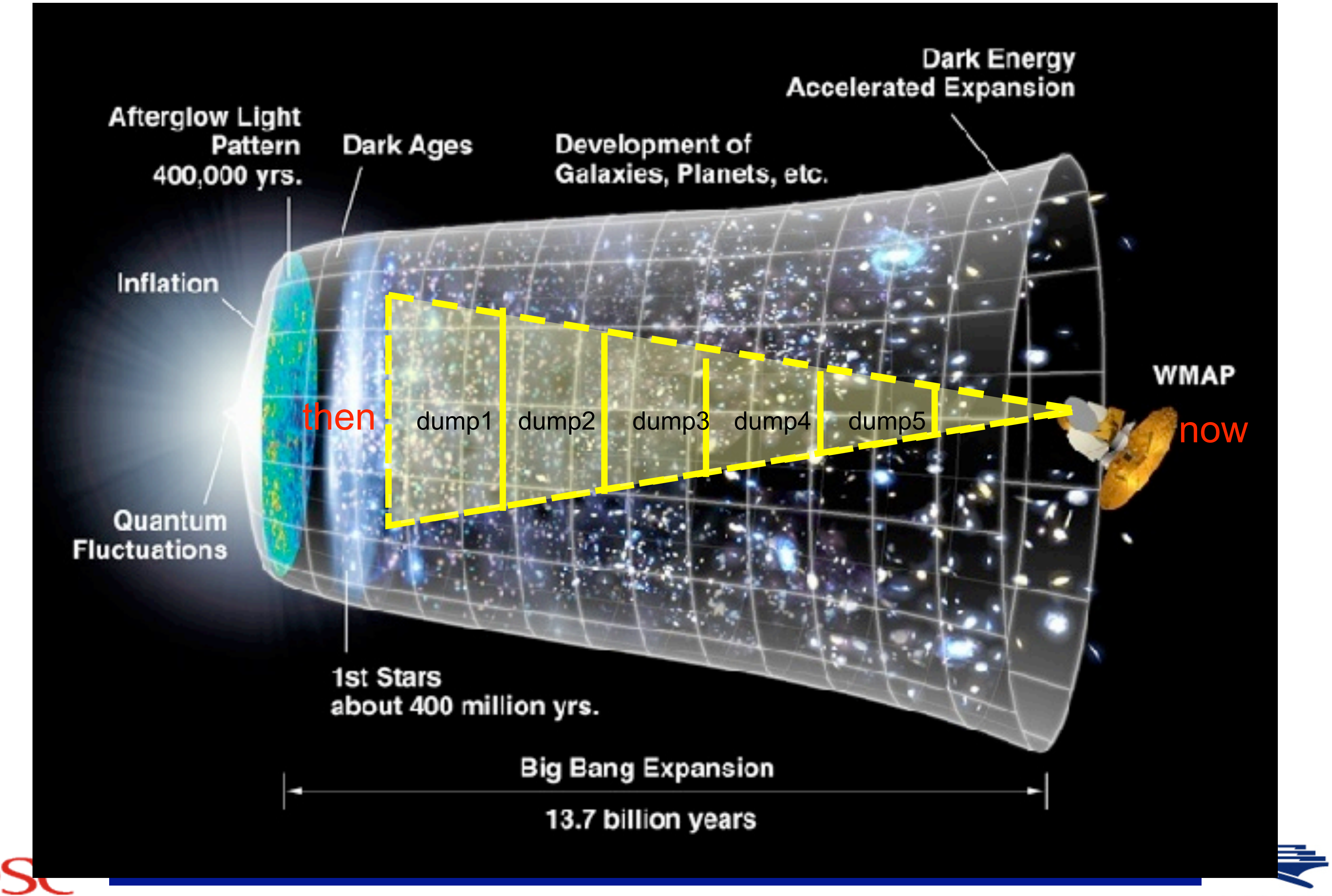
	C	A	G	C	C	U	C	G	C	U	U	A	G
A	0,0	1,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	1,0	0,0
A	0,0	1,0	0,7	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	1,0	0,7
U	0,0	0,0	0,8	0,3	0,0	0,0	0,0	0,0	0,0	1,0	1,0	0,0	0,7
G	0,0	0,0	1,0	0,3	0,0	0,0	0,7	1,0	0,0	0,0	0,7	0,7	1,0
C	1,0	0,0	0,0	2,0	1,3	0,3	1,0	0,3	2,0	0,7	0,3	0,3	0,3
C	1,0	0,7	0,0	1,0	3,0	1,7	?						
A													
U													
U													
G													
A													
C													
G													
G													

Next Generation Cosmology



Evolving the Universe from the "Big Bang"

Composing simulation outputs from different timeframes builds up light-cone volume

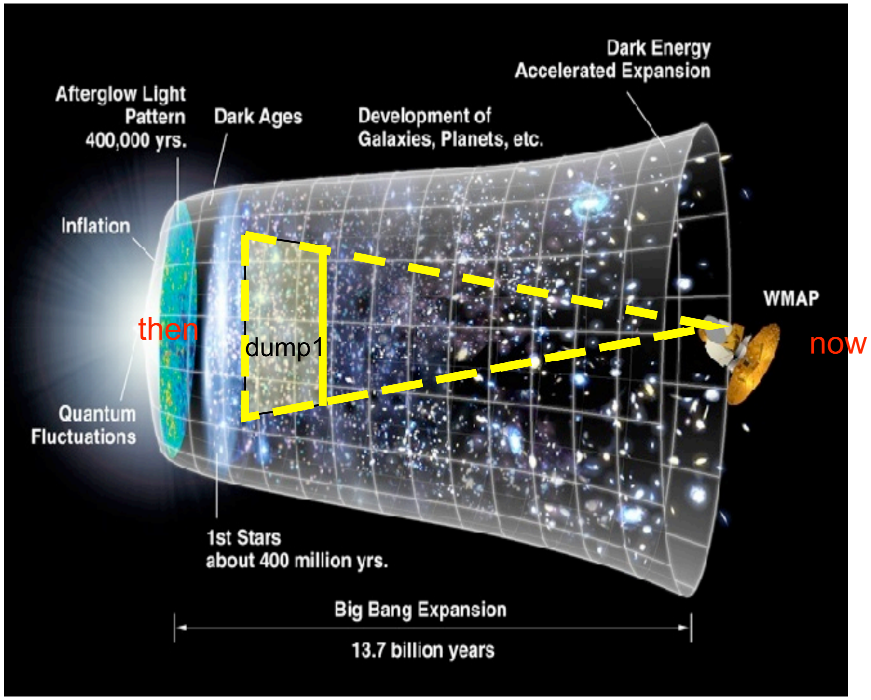

Afterglow Light Pattern 400,000 yrs.
Dark Ages
Development of Galaxies, Planets, etc.
Dark Energy Accelerated Expansion
Inflation
then
dump1
dump2
dump3
dump4
dump5
WMAP
now
Quantum Fluctuations
1st Stars about 400 million yrs.
Big Bang Expansion
13.7 billion years

Slide modified from Mike Norman

The Universe's First Billion Years after the "Big Bang"

- **ENZO** simulates the first billion years of cosmic evolution after the "Big Bang"

- Key period which represents
  - A tumultuous period of intense star formation *throughout the universe*
  - Synthesis of the first heavy elements in massive stars
  - Supernovae, gamma-ray bursts, seed black holes, and the corresponding growth of supermassive black holes and the birth of quasars
  - Assembly of first galaxies



Afterglow Light Pattern 400,000 yrs.
Dark Ages
Development of Galaxies, Planets, etc.
Dark Energy Accelerated Expansion
Inflation
then
dump1
WMAP
now
Quantum Fluctuations
1st Stars about 400 million yrs.
Big Bang Expansion
13.7 billion years

SAN DIEGO SUPERCOMPUTER CENTER

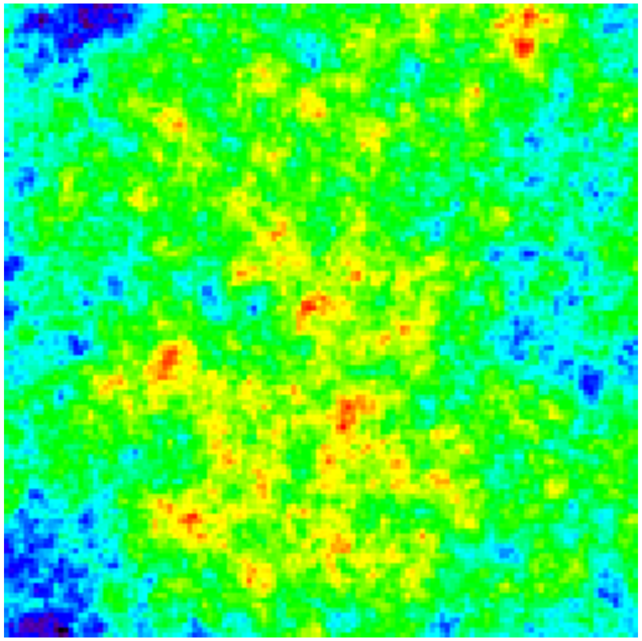
Slide modified from Mike Norman

UCSD

ENZO Simulations

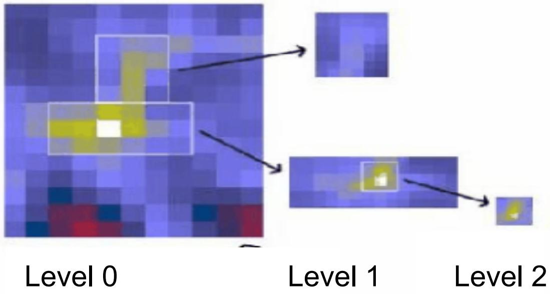
What ENZO does:

- Calculates the growth of cosmic structure from seed perturbations to form stars, galaxies, and galaxy clusters, including simulation of
  - *Dark matter*
  - *Ordinary matter (atoms)*
  - *Self-gravity*
  - *Cosmic expansion*
- Uses adaptive mesh refinement (AMR) to provide high spatial resolution in 3D
  - The Santa Fe light cone simulation generated over 350,000 grids at 7 levels of refinement
  - **Effective resolution = $65{,}536^3$**



Formation of a galaxy cluster

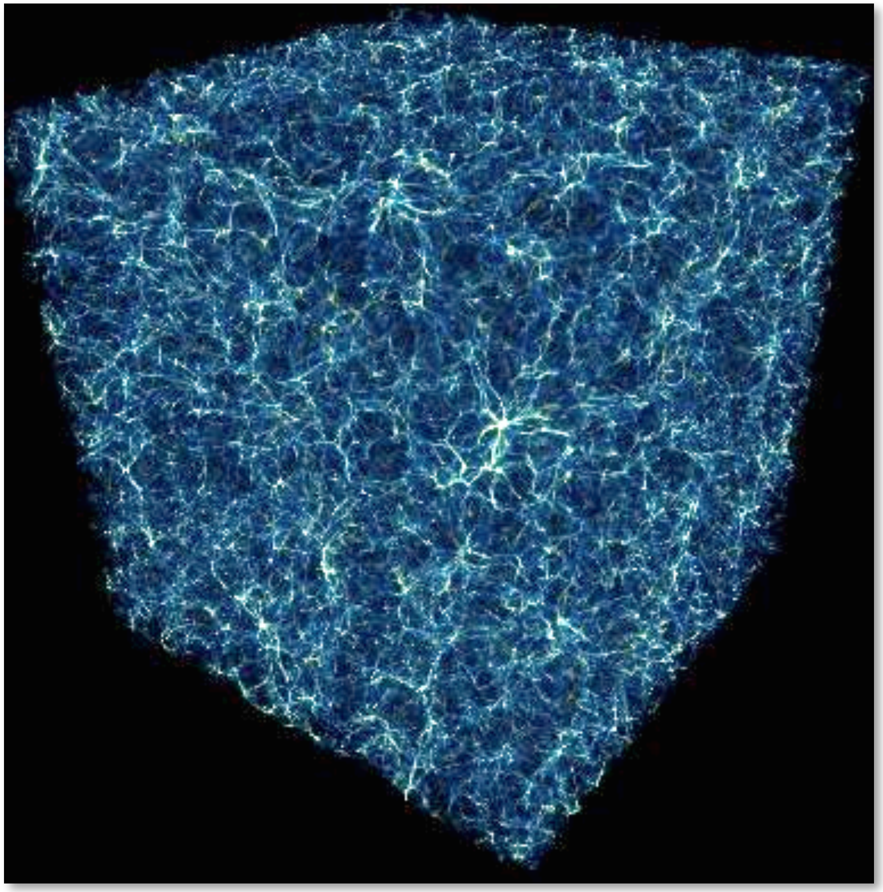
AMR



Slide modified from Mike Norman

Enzo Data Volumes

- 2048^3 simulation (2008)
  - 8 gigazones X
  - 16 fields/zone X
  - 4 bytes/field
  - = 0.5 TB/output X
  - 100 outputs/run
  - = **50 TB**
- 4096^3 simulation (2009)
  - 64 gigazones X
  - 16 fields/zone X
  - 4 bytes/field
  - = 4 TB/output X
  - 50 outputs/run
  - = **200 TB**



Greater Simulation Accuracy Requires More Computing and Generates More Data

ENZO at Petascale (10^15)

- Self-consistent **radiation-**hydro simulations of structural, chemical, and radiative evolution of the universe simulates from first stars to first galaxies

Computer Science challenges

- Parallelizing the grid hierarchy metadata for millions of subgrids distributed across 10s of thousands of cores
- Efficient dynamic load balancing of the numerical computations, taking memory hierarchy and latencies into account
- Efficient parallel "packed AMR" I/O for 100 TB data dumps
- Inline data analysis/viz. to reduce I/O

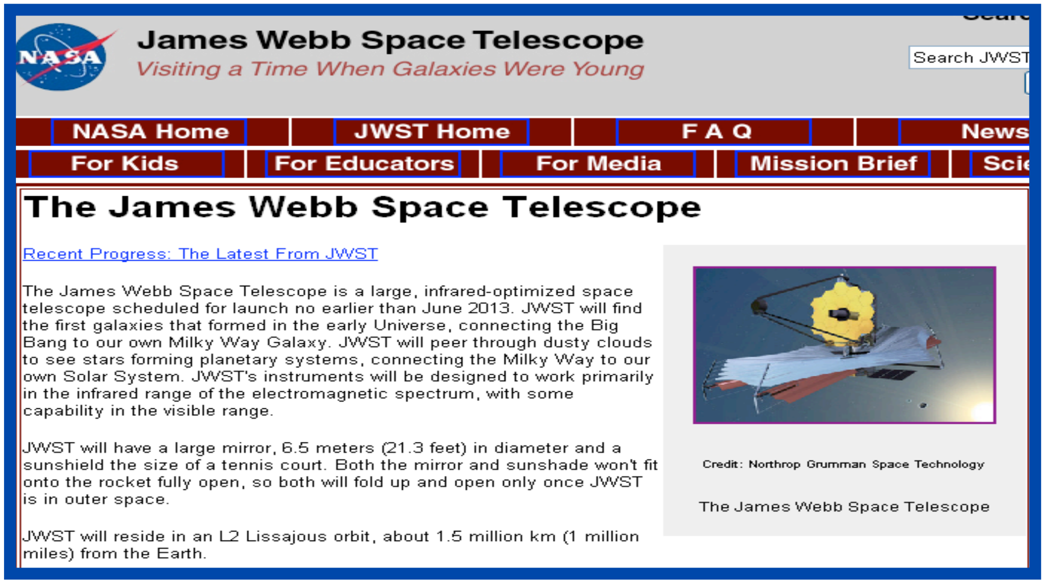
Slide modified from Mike Norman, image by Robert Harkness

Verifying Theory with Observation

- **James Webb Space Telescope**, coming in 2013 will probe the first billion years of the universe – providing observations of unprecedented depth and breadth

- Data will enable tight integration of observation and theory, and will enable simulations to approach realistic complexity

- Analysis of **petascale data sets** will be essential for validating model



James Webb Space Telescope

Visiting a Time When Galaxies Were Young

The James Webb Space Telescope

Recent Progress: The Latest From JWST

The James Webb Space Telescope is a large, infrared-optimized space telescope scheduled for launch no earlier than June 2013. JWST will find the first galaxies that formed in the early Universe, connecting the Big Bang to our own Milky Way Galaxy. JWST will peer through dusty clouds to see stars forming planetary systems, connecting the Milky Way to our own Solar System. JWST's instruments will be designed to work primarily in the infrared range of the electromagnetic spectrum, with some capability in the visible range.

JWST will have a large mirror, 6.5 meters (21.3 feet) in diameter and a sunshield the size of a tennis court. Both the mirror and sunshade won't fit onto the rocket fully open, so both will fold up and open only once JWST is in outer space.

JWST will reside in an L2 Lissajous orbit, about 1.5 million km (1 million miles) from the Earth.

Credit: Northrop Grumman Space Technology

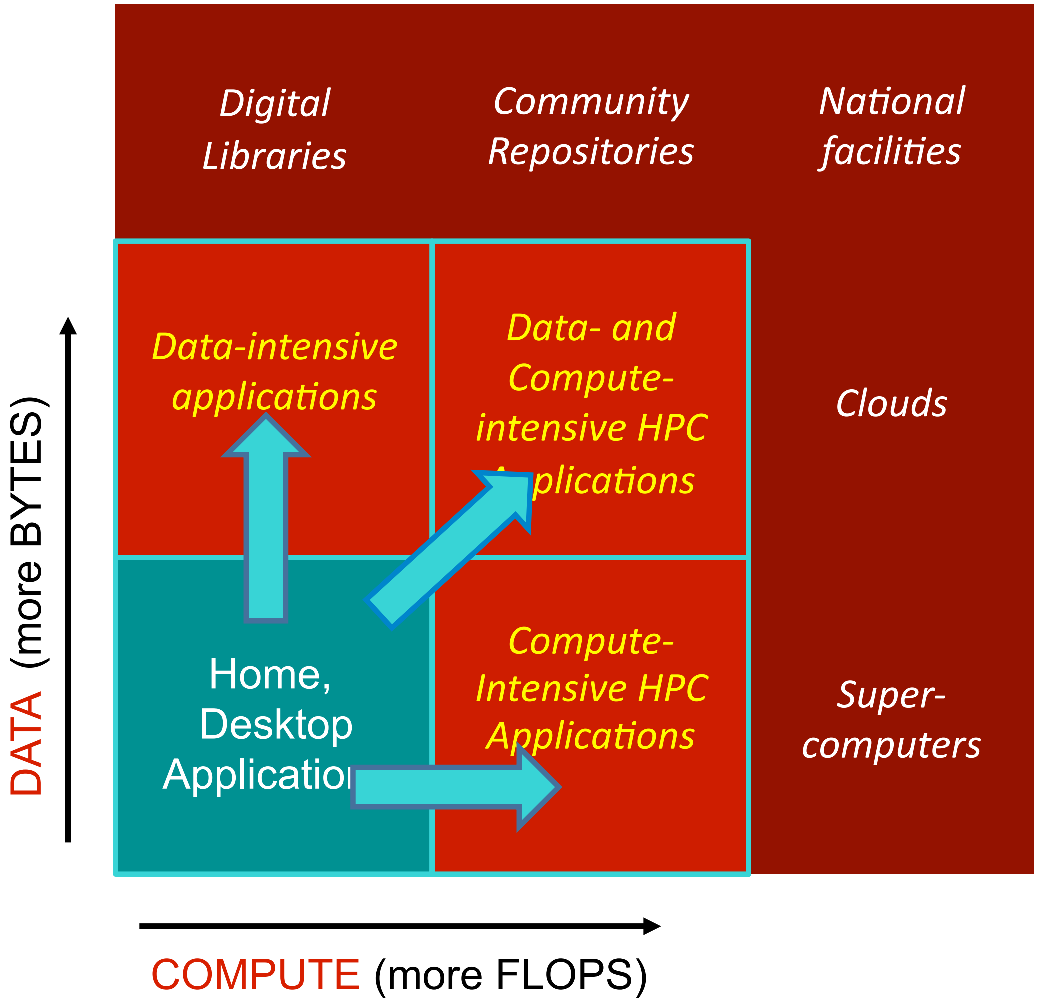
The James Webb Space Telescope

Slide modified from Mike Norman, image by Robert Harkness

SDSC Cyberinfrastructure Resources and Services

Enabling Ideas

The focus of SDSC's resources and services is to provide an integrated environment that empowers users to move beyond their local boundaries to further research and education goals



Digital Libraries

Community Repositories

National facilities

Data-intensive applications

Data- and Compute-intensive HPC Applications

Clouds

Home, Desktop Applications

Compute-Intensive HPC Applications

Super-computers

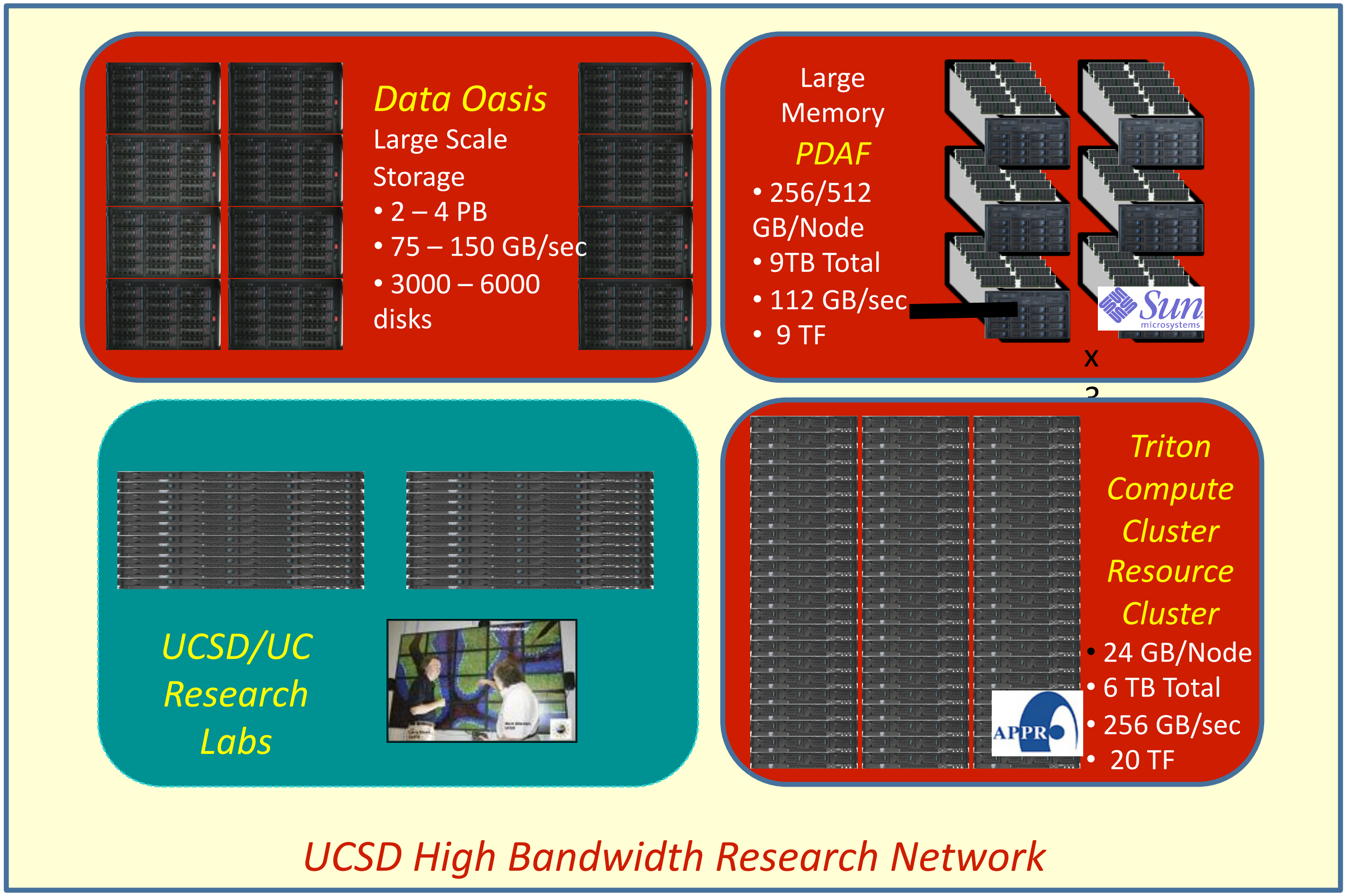
DATA (more BYTES)

COMPUTE (more FLOPS)

Triton
Resource

Triton Resource Draft Configuration

Will go into Production by Summer 2009



Data Oasis
Large Scale Storage
- 2 – 4 PB
- 75 – 150 GB/sec
- 3000 – 6000 disks

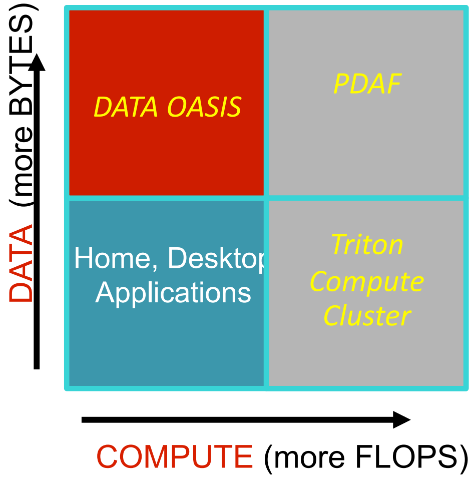
Large Memory
PDAF
- 256/512 GB/Node
- 9TB Total
- 112 GB/sec
- 9 TF

x 3

UCSD/UC Research Labs

Triton Compute Cluster Resource Cluster
- 24 GB/Node
- 6 TB Total
- 256 GB/sec
- 20 TF

UCSD High Bandwidth Research Network



DATA (more BYTES)

DATA OASIS

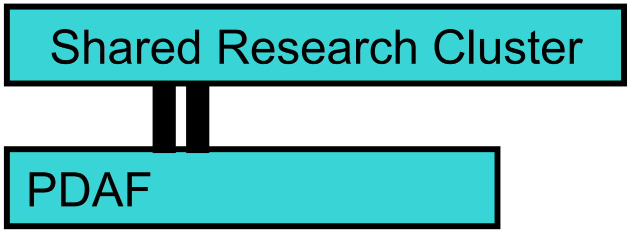
PDAF

Home, Desktop Applications

Triton Compute Cluster

COMPUTE (more FLOPS)

The Triton Resource: Data Oasis



Shared Research Cluster

PDAF

Data Oasis
High Performance
Disk Storage

- Integrated
- Expandable
- Efficient

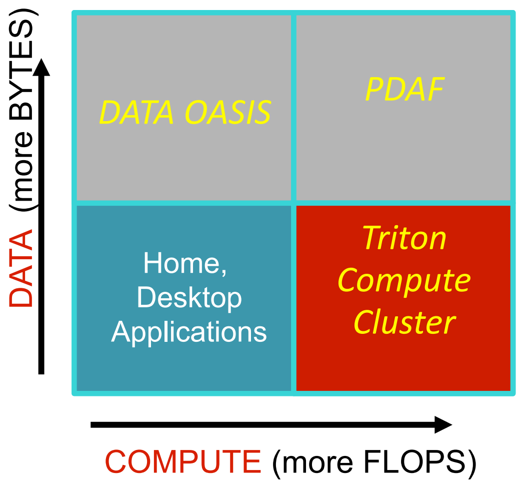
- Data Oasis will be used as storage for
  - PDAF, Triton Compute Cluster, UCSD black boxes, in situ instruments and clusters, UC colleagues, key projects

- Administered by SDSC and the UCSD Libraries

What will we do with the Data Oasis?

- **Faculty Terabytes** – Archival disk storage for UCSD faculty data
- **Storage Services:** Backups, DB hosting, Project leverage
- **Research Data Services:** Data visualization, database development, portals, GIS, data mining, statistical analysis, etc.
- **Chronopolis** preservation services

Triton's Expandable Triton Compute Cluster



DATA (more BYTES)

DATA OASIS	*PDAF*
Home, Desktop Applications	*Triton Compute Cluster*

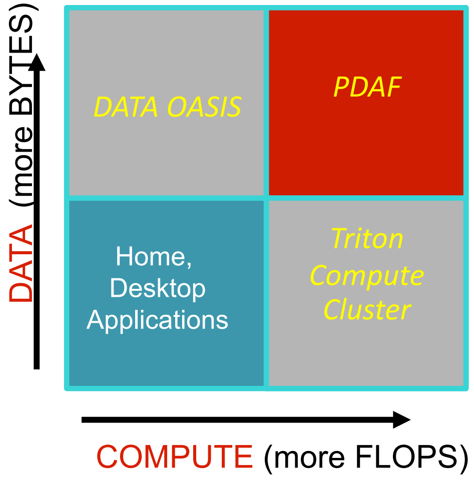
COMPUTE (more FLOPS)

- Triton Triton Compute Cluster is a launch system that provides a seed for expandable "condo-style" core computational facilities
- Full suite of centralized support services (24X7 operations, *security, networking, SW,* storage, support, training, documentation, system administration)

"Condo" facilities are green and expandable

- Economies of scale for capital purchase and operations
- Whole is greater than the sum of small parts – can aggregate across machine for "capability" runs during idle periods

The Triton Resource: PDAF



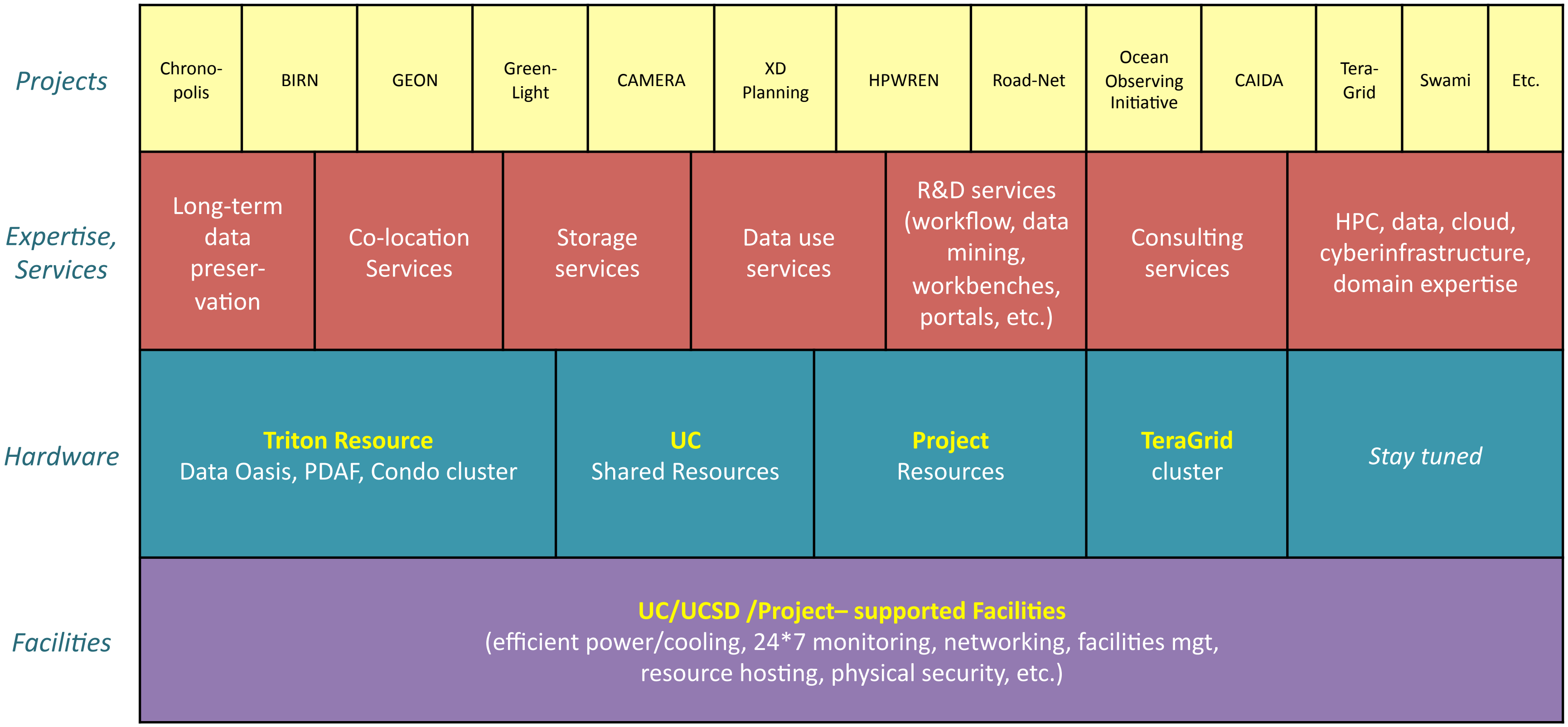
Triton Resource

- Petascale Data Analysis Facility supports SDSC Research focus
- Campus planning group of users determined practical balancing (compute, memory, storage, network connectivity) of Triton
- Configured for analysis, modeling, simulation of *very* large data sets

PDAF overview:

- Large-Memory Data-intensive cluster
- 8 x 512GB + 20 x 256 GB nodes
  - 8 AMD 8380 Shanghai Processors at 2.5GHz. 4 X 10Gbit Myrinet.
  - Sun x4600M2
  - 4 Nodes will have large local storage for DBs (2 @ 7.6TB, 2@ 24TB)
  - 9 TB Total
- Linux based
- Energy efficient
- High speed network to Data Oasis

SDSC Overall "Architecture"



Projects

Chrono-polis | BIRN | GEON | Green-Light | CAMERA | XD Planning | HPWREN | Road-Net | Ocean Observing Initiative | CAIDA | Tera-Grid | Swami | Etc.

Expertise, Services

Long-term data preservation | Co-location Services | Storage services | Data use services | R&D services (workflow, data mining, workbenches, portals, etc.) | Consulting services | HPC, data, cloud, cyberinfrastructure, domain expertise

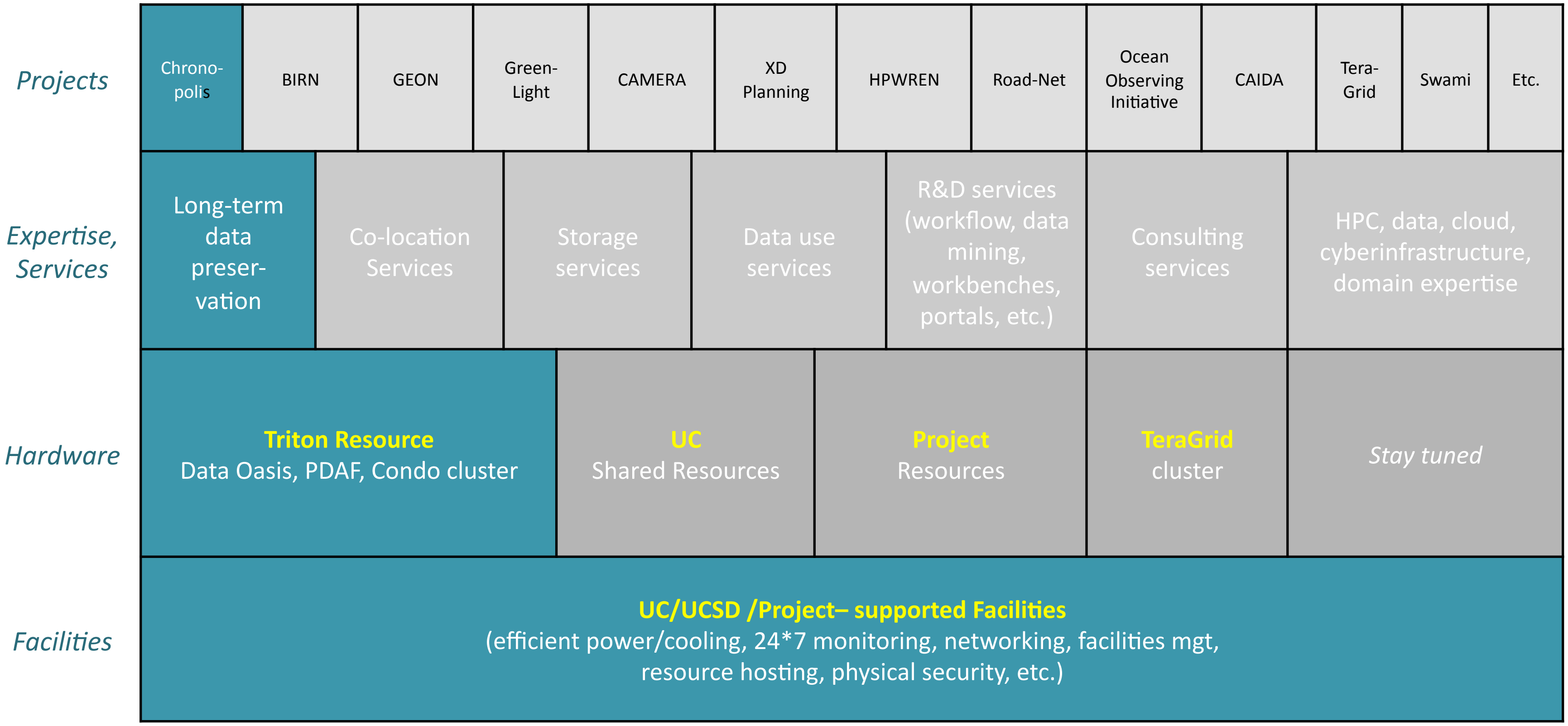
Hardware

Triton Resource Data Oasis, PDAF, Condo cluster | **UC** Shared Resources | **Project** Resources | **TeraGrid** cluster | *Stay tuned*

Facilities

UC/UCSD /Project– supported Facilities
(efficient power/cooling, 24*7 monitoring, networking, facilities mgt, resource hosting, physical security, etc.)

Data Cyberinfrastructure for
Long-term Preservation



Projects

Chronopolis | BIRN | GEON | GreenLight | CAMERA | XD Planning | HPWREN | Road-Net | Ocean Observing Initiative | CAIDA | TeraGrid | Swami | Etc.

Expertise, Services

Long-term data preservation | Co-location Services | Storage services | Data use services | R&D services (workflow, data mining, workbenches, portals, etc.) | Consulting services | HPC, data, cloud, cyberinfrastructure, domain expertise

Hardware

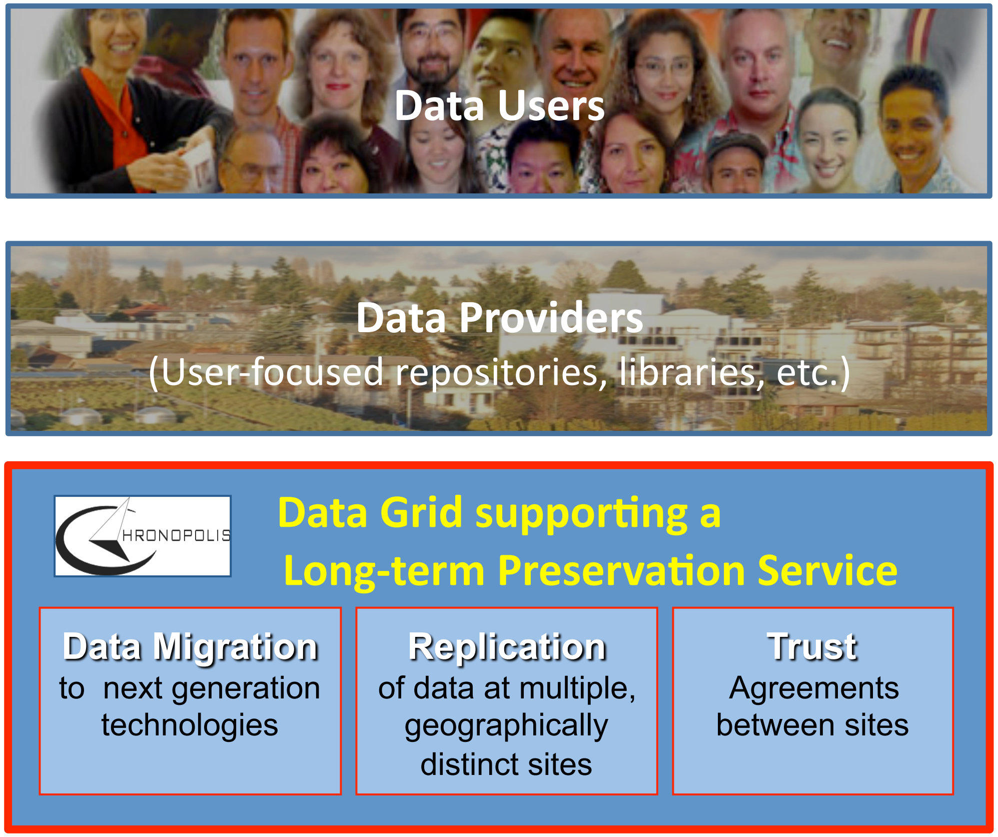
Triton Resource Data Oasis, PDAF, Condo cluster | **UC** Shared Resources | **Project** Resources | **TeraGrid** cluster | *Stay tuned*

Facilities

UC/UCSD /Project– supported Facilities
(efficient power/cooling, 24*7 monitoring, networking, facilities mgt, resource hosting, physical security, etc.)

Long-term Preservation Service and Resources

What is Chronopolis?


Data Users
Data Providers
(User-focused repositories, libraries, etc.)
HRONOPOLIS
Data Grid supporting a
Long-term Preservation Service
Data Migration
to next generation technologies
Replication
of data at multiple, geographically distinct sites
Trust
Agreements between sites

Who is Chronopolis?

Chronopolis is being developed by a national consortium led by **SDSC and the UCSD Libraries (UCSDL)** and funded by the **Library of Congress**.

Initial Chronopolis nodes include:

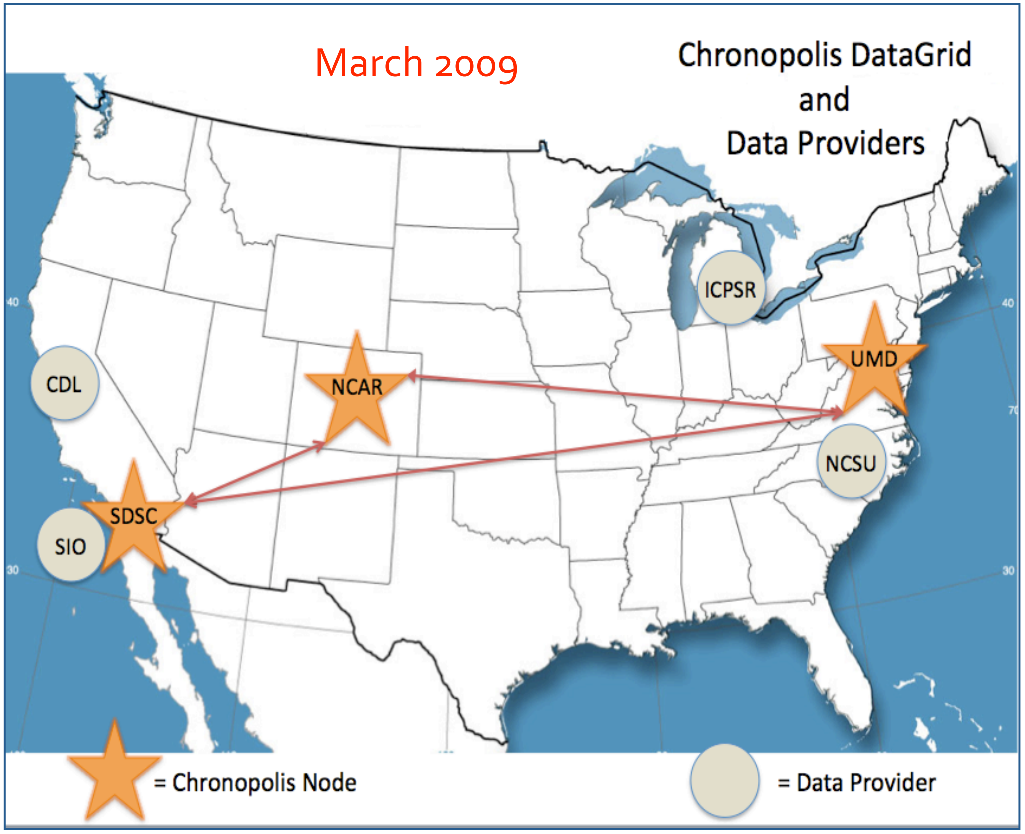
- *SDSC/UCSDL at UCSD*
- *University of Maryland Institute for Advanced Computer Studies (UMIACS)*
- *National Center for Atmospheric Research (NCAR) in Boulder, CO*

Current Chronopolis Collections

http://chronopolis.sdsc.edu

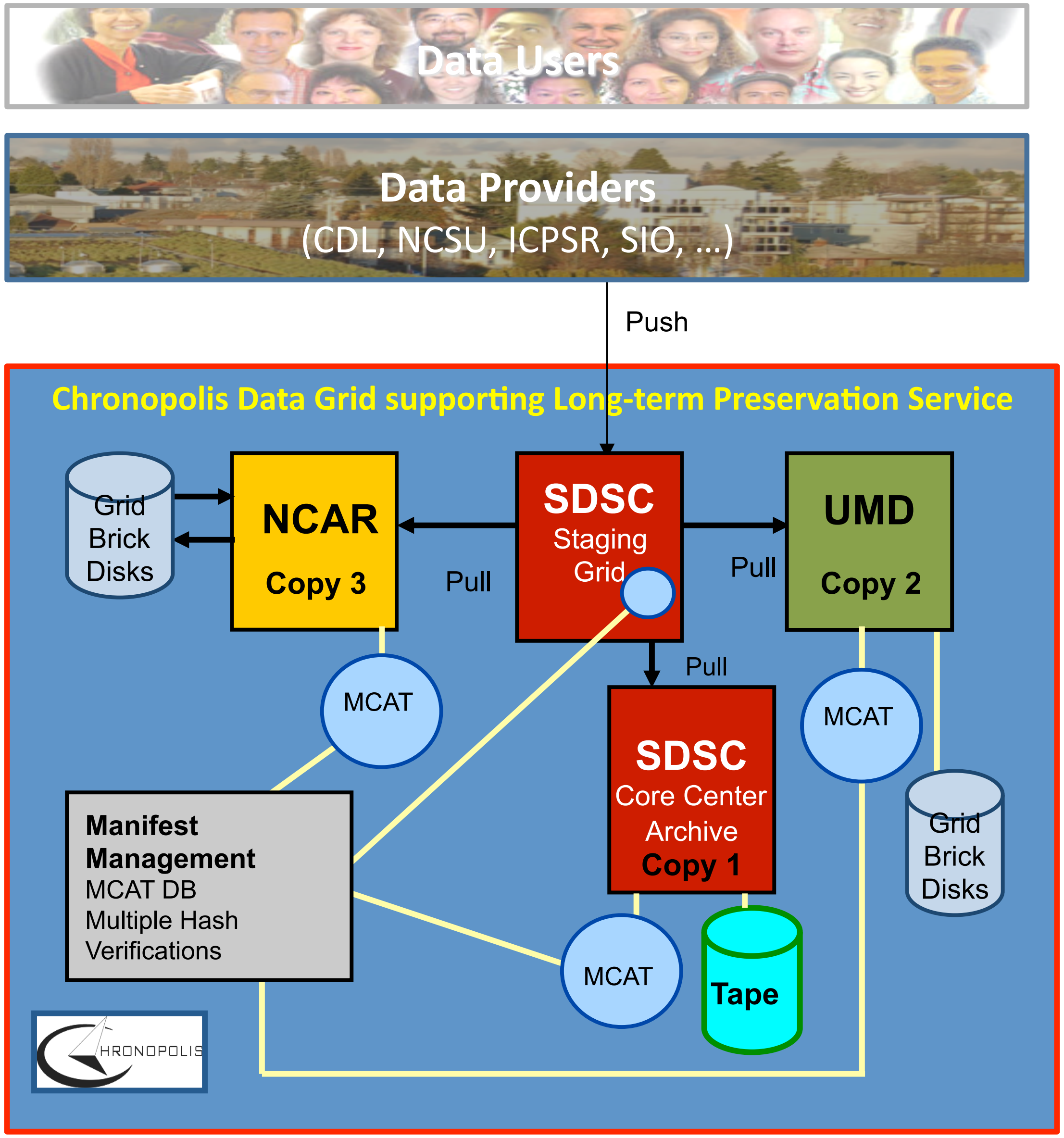
Collections from Data Providers:

- **Inter-university Consortium of Political and Social Research** – preservation copy of all collections including 40 years of social science data and Census (8 TB)
- **California Digital Library** – political and government web crawls, Web-at-risk collection (5 TB)
- **SIO Explorer** – data from 50 years of research voyages (1 TB)
- **NCSU Libraries** -- State and local geospatial data (6 TB)



Inside Chronopolis

- Sites linked by main staging grid where data is verified for integrity, and quarantined for security purposes.
- Collections independently pulled into each system.
- Manifest layer provides added security for database management and data integrity validation.
- Benefits
  - Each collection copy independently managed
  - Collections available from each site?
  - High reliability



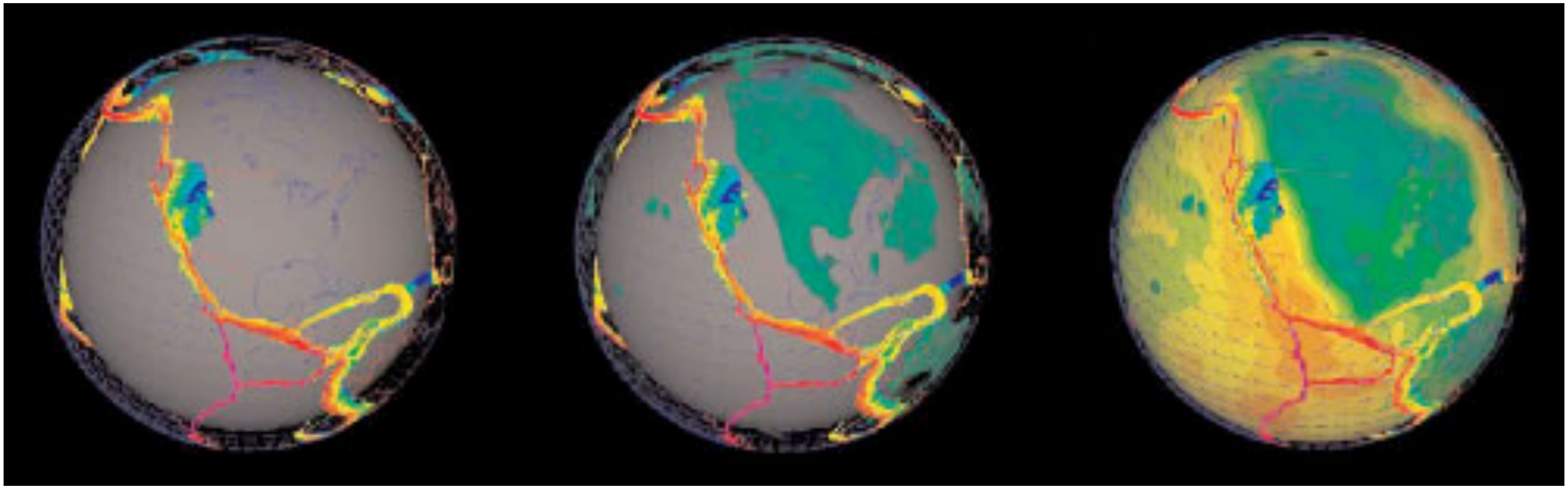
SDSC and the UCSD Libraries – a Unique Organizational Relationship

- SDSC and the UCSD Libraries working together to support the entire data life cycle – from curation and ingest to storage and preservation

- Joint SDSC/UCSDL team has worked with National Science Foundation, NARA, and the U.S. Library of Congress to pioneer integrated approaches to digital data cyberinfrastructure

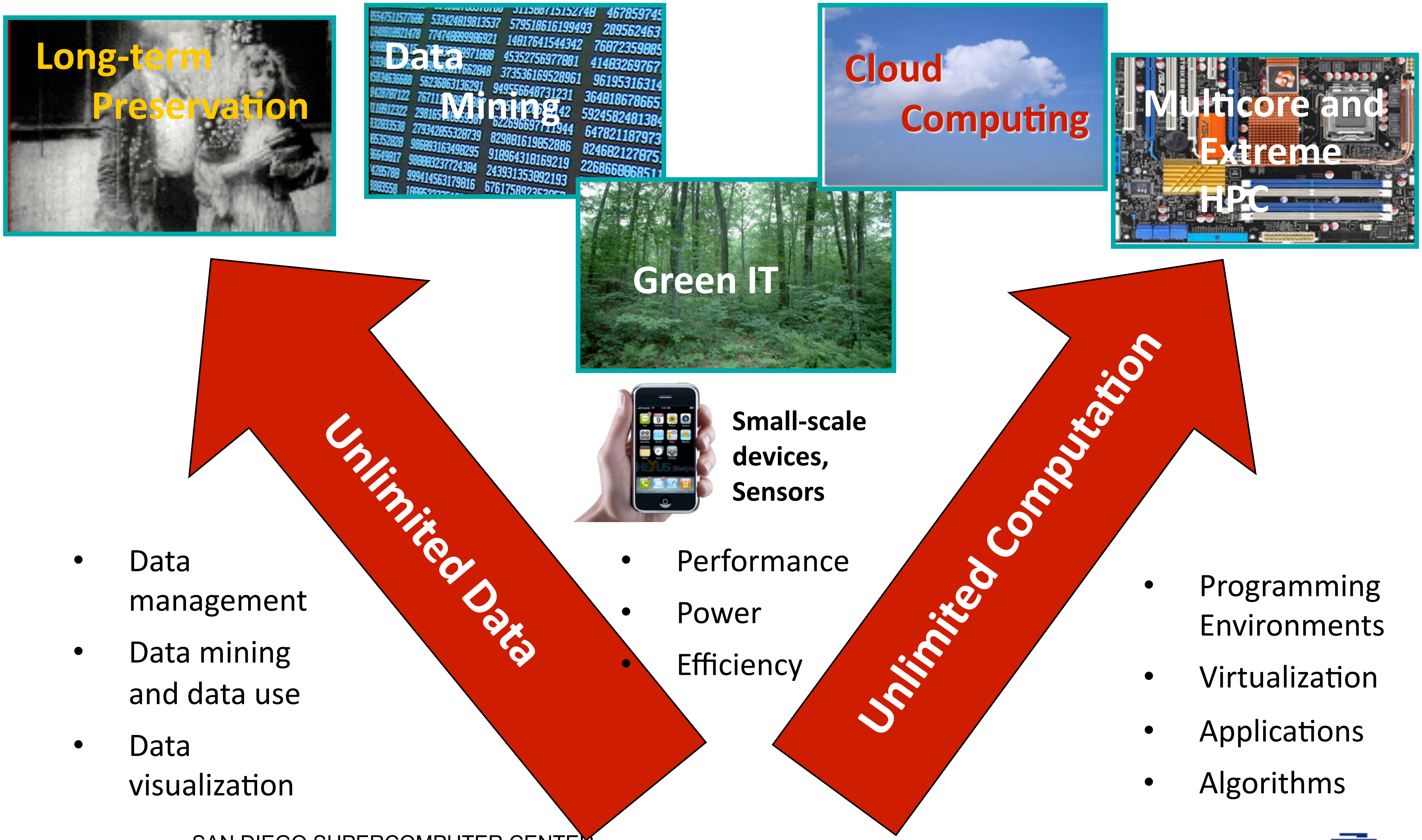
In addition to Chronopolis, SDSC/ UCSDL worked with the ***Library of Congress*** *on a pilot to develop* ***distributed preservation solutions*** *for valued collections like the Prokudin-Gorskii photographs of the Russian Revolution*

Cyberinfrastructure for the Next Decade



Fran Berman

Significant Trends in Cyberinfrastructure



Small-scale devices, Sensors

- Data management
- Data mining and data use
- Data visualization

- Performance
- Power
- Efficiency

- Programming Environments
- Virtualization
- Applications
- Algorithms

The Next Decade will see Increased Constraints

amazon webservices™

Eucalyptus

Economics

Cyberinfrastructure will need to support both a broader set of use cases and a greater set of limiting factors

Regulation, Policy

HEALTH INSURANCE PORTABILITY and ACCOUNTABILITY ACT

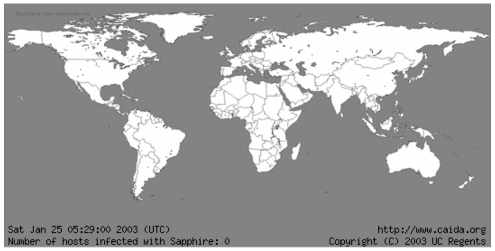
HIPAA

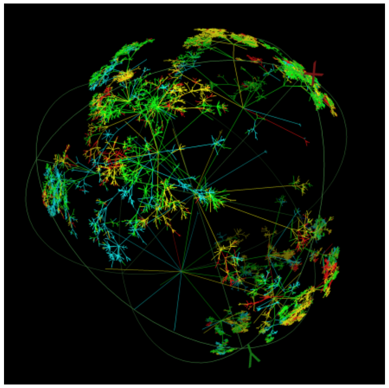
ADMINISTRATIVE SIMPLIFICATION: PRIVACY, SECURITY, TRANSACTIONS

Power

Low-cost physical environments for Cyberinfrastructure have become increasingly important as systems scale

The Next Decade will see Increased Complexity


Sat Jan 25 05:29:00 2003 (UTC)
Number of hosts infected with Sapphire: 0
http://www.caida.org
Copyright (C) 2003 UC Regents



Scale

Globalization

UniProt
the universal protein resource

Interoperability

PDB
PROTEIN DATA BANK

More advanced applications and more capable systems will require an unprecedented degree of integration, interoperability, coordination, and protections

The Next Decade will see Increased Opportunity

amazon.com

China People Place Culture History (Hardcover)

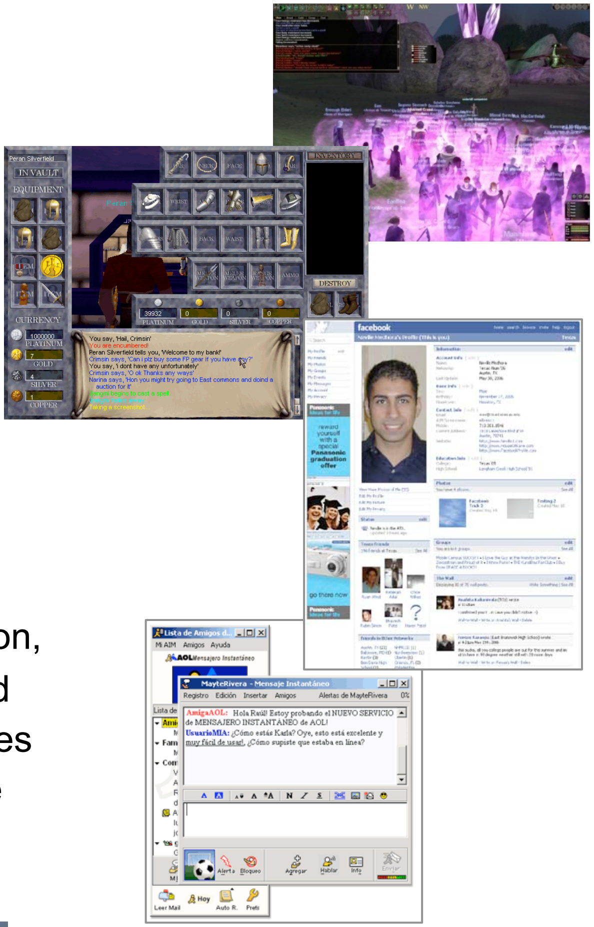
In Stock

Check Out Related Media

Special Offers and Product Promotions

Frequently Bought Together

Data → Information → Knowledge

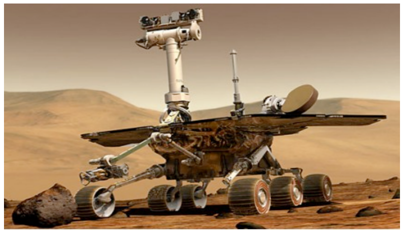


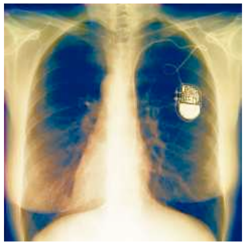
You say, 'Hail, Crimsin'
You are encumbered!
Peran Silverfield tells you, 'Welcome to my bank!'
Crimsin says, 'Can i plz buy some FP gear if you have any?'
You say, 'i dont have any unfortunately'
Crimsin says, 'O ok Thanks any ways'
Narina says, 'Hon you might try going to East commons and doind a auction for it'
Jjangmi begins to cast a spell.
Jjangmi fades away.
Taking a screenshot...

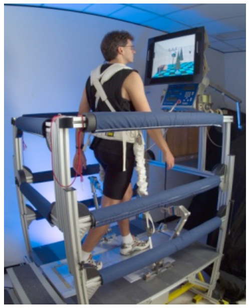
AmigaAOL: Hola Raúl! Estoy probando el NUEVO SERVICIO de MENSAJERO INSTANTÁNEO de AOL!
UsuarioMIA: ¿Cómo estás Karla? Oye, esto está excelente y muy fácil de usar!, ¿Cómo supiste que estaba en línea?

Boundaries between domains are blurring, providing greater opportunities for innovation, agility, and synergy, and presenting new challenges for Cyberinfrastructure

Cyber-physical systems







Social networks, Mass communication

SDSC Focus will be to Continue Working with Researchers at the Cutting Edge

PACIFIC RIM APPLICATIONS AND GRID MIDDLEWARE ASSEMBLY
PRAGMA

Kepler

ROCKS

PDB
PROTEIN DATA BANK

SIOExplorer

American Red Cross

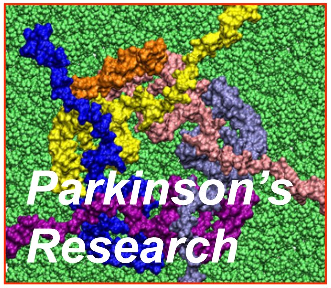
Safe and Well List

CIPRES

CAIDA

Next Generation Biology Workbench

TeraGrid™


Parkinson's Research

PERI

Thank you

www.sdsc.edu