cni
Coalition for Networked Information
Fall Task Force Meeting
Dec. 10-11, 2007

DATA CYBERINFRASTRUCTURE COLLABORATION at the University of California, San Diego

Brian E. C. Schottlaender
University Librarian – University of California, San Diego

Robert H. McDonald
Director, Strategic Data Alliances – San Diego Supercomputer Center

Outline

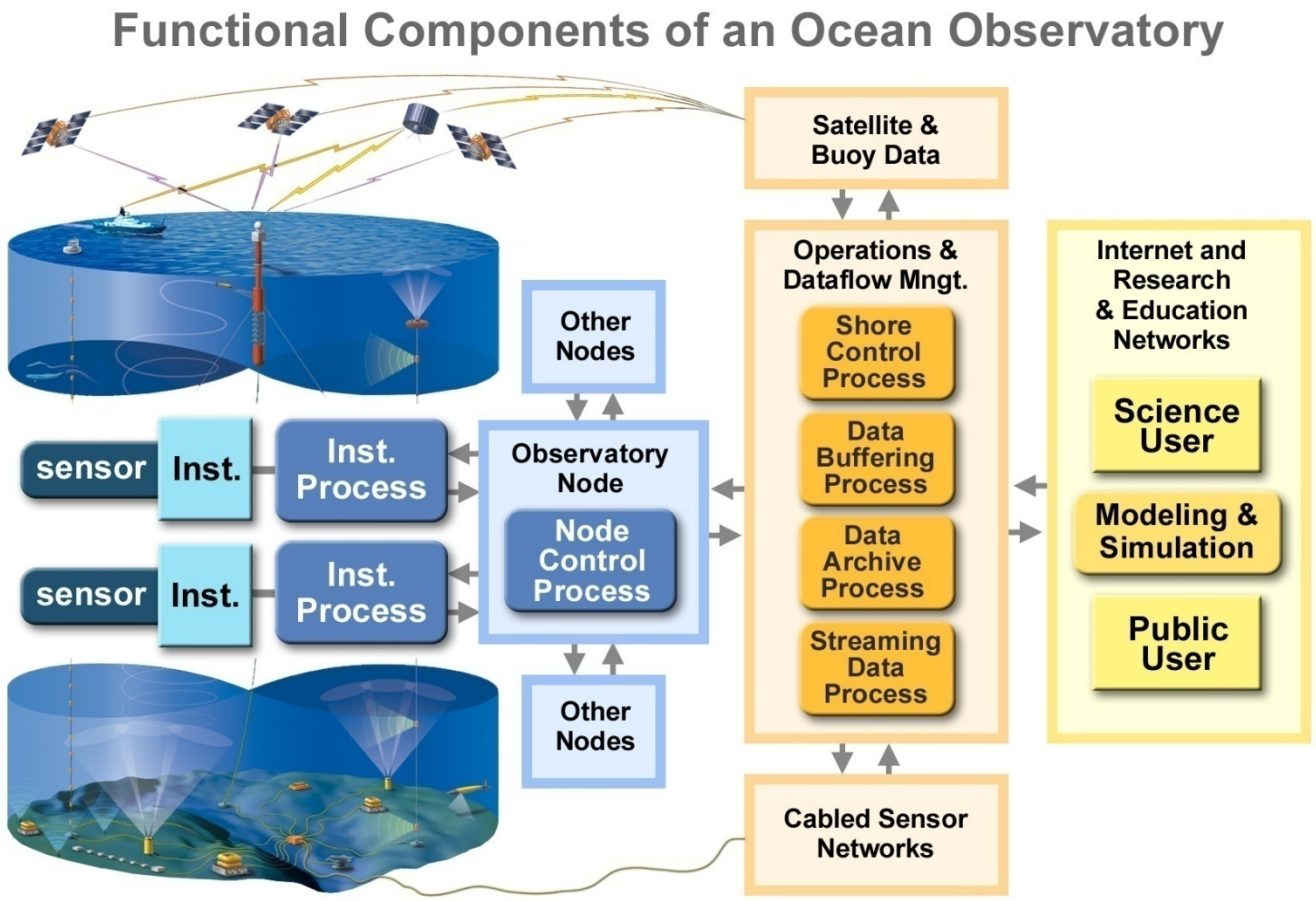
- What is cyberinfrastructure?
- Collaboration: Turning need into opportunity
- The technology at play at local, system, and national levels
- Lessons learned
- Future plans

What Is Cyberinfrastructure?

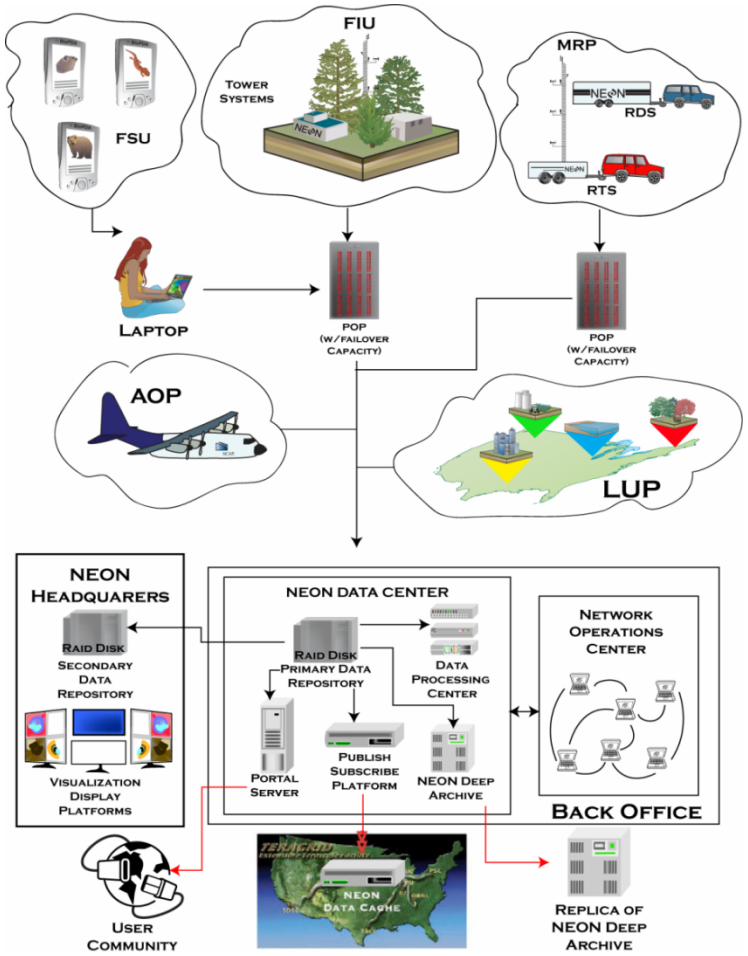
Cyberinfrastruture is the coordinated aggregate of software, hardware and other technologies, as well as human expertise, required to support current and future discoveries in science and engineering.

— Fran Berman, Director
San Diego Supercomputer Center (SDSC)

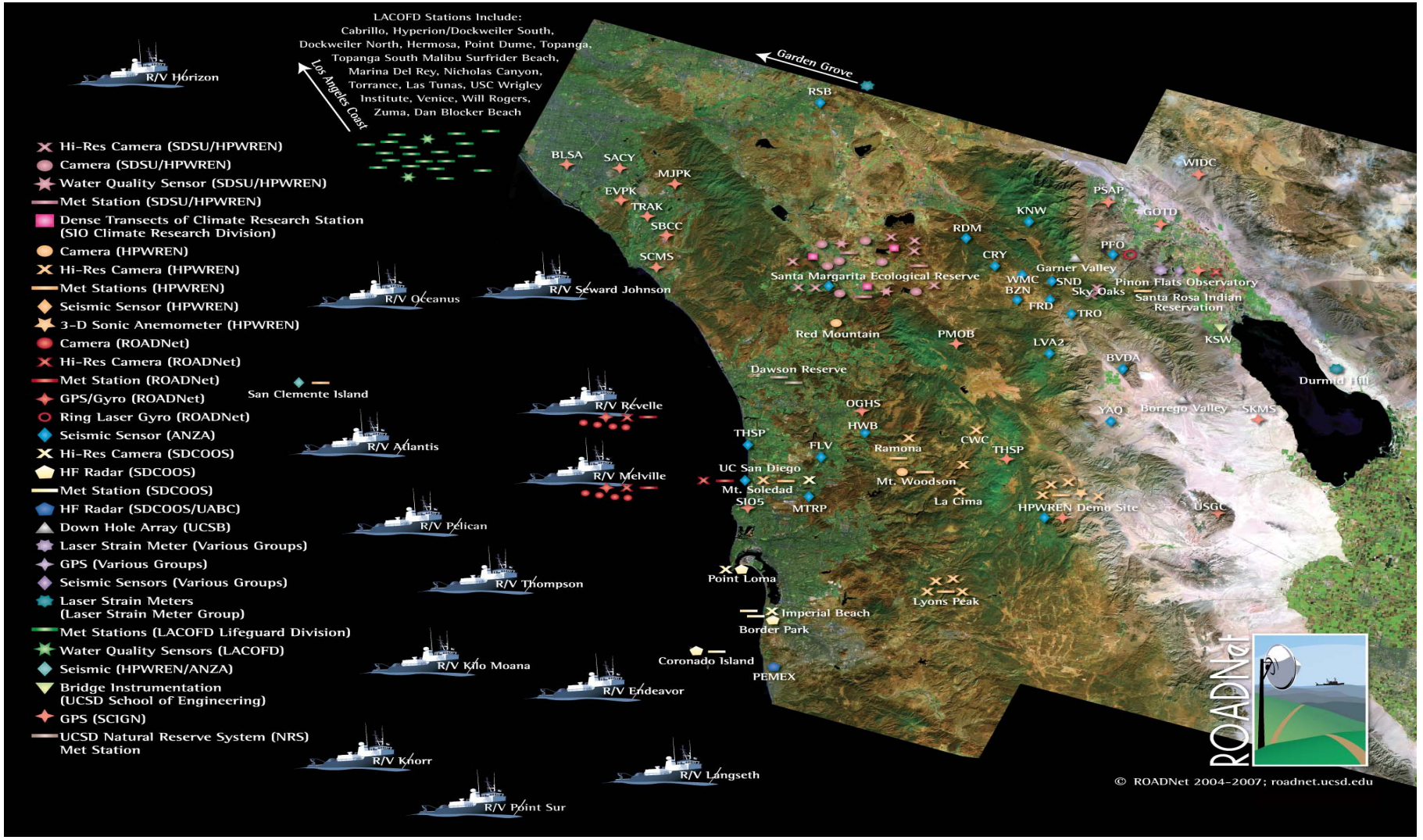
Ocean Observatories Initiative (OOI)



National Ecological Observatory Network (Neon)



ROADNet


LACOFD Stations Include:
Cabrillo, Hyperion/Dockweiler South, Dockweiler North, Hermosa, Point Dume, Topanga, Topanga South Malibu Surfrider Beach, Marina Del Rey, Nicholas Canyon, Torrance, Las Tunas, USC Wrigley Institute, Venice, Will Rogers, Zuma, Dan Blocker Beach
Los Angeles Coast
Garden Grove
R/V Horizon
Hi-Res Camera (SDSU/HPWREN)
Camera (SDSU/HPWREN)
Water Quality Sensor (SDSU/HPWREN)
Met Station (SDSU/HPWREN)
Dense Transects of Climate Research Station (SIO Climate Research Division)
Camera (HPWREN)
Hi-Res Camera (HPWREN)
Met Stations (HPWREN)
Seismic Sensor (HPWREN)
3-D Sonic Anemometer (HPWREN)
Camera (ROADNet)
Hi-Res Camera (ROADNet)
Met Station (ROADNet)
GPS/Gyro (ROADNet)
Ring Laser Gyro (ROADNet)
Seismic Sensor (ANZA)
Hi-Res Camera (SDCOOS)
HF Radar (SDCOOS)
Met Station (SDCOOS)
HF Radar (SDCOOS/UABC)
Down Hole Array (UCSB)
Laser Strain Meter (Various Groups)
GPS (Various Groups)
Seismic Sensors (Various Groups)
Laser Strain Meters (Laser Strain Meter Group)
Met Stations (LACOFD Lifeguard Division)
Water Quality Sensors (LACOFD)
Seismic (HPWREN/ANZA)
Bridge Instrumentation (UCSD School of Engineering)
GPS (SCIGN)
UCSD Natural Reserve System (NRS) Met Station
R/V Oceanus
R/V Seward Johnson
San Clemente Island
R/V Revelle
R/V Atlantis
R/V Melville
R/V Pelican
R/V Thompson
R/V Kilo Moana
R/V Endeavor
R/V Knorr
R/V Point Sur
R/V Langseth
RSB
BLSA
SACY
MJPK
EVPK
TRAK
SBCC
SCMS
Santa Margarita Ecological Reserve
RDM
KNW
WIDC
PSAP
GOTD
PFO
CRY
Garner Valley
WMC
SND
BZN
Sky Oaks
Pinon Flats Observatory
Santa Rosa Indian Reservation
FRD
TRO
Red Mountain
PMOB
LVA2
KSW
BVDA
Dawson Reserve
Durmid Hill
OGHS
HWB
YAQ
Borrego Valley
SKMS
THSP
FLV
Ramona
CWC
THSP
UC San Diego
Mt. Soledad
SIO5
MTRP
Mt. Woodson
La Cima
HPWREN Demo Site
USGC
Point Loma
Lyons Peak
Imperial Beach
Border Park
Coronado Island
PEMEX
ROADNet
© ROADNet 2004-2007; roadnet.ucsd.edu

Common Characteristic

Data

Lots of data

Petabytes of data

Examples

UCSD Libraries DAMS → 6 TB (800K Images)

CDL → ramping up to 40 TB

Chronopolis → +50 TB per year (2007-2011)

Sloan Digital Sky Survey → 40 TB every 3 days

LSST → +10 TB per night

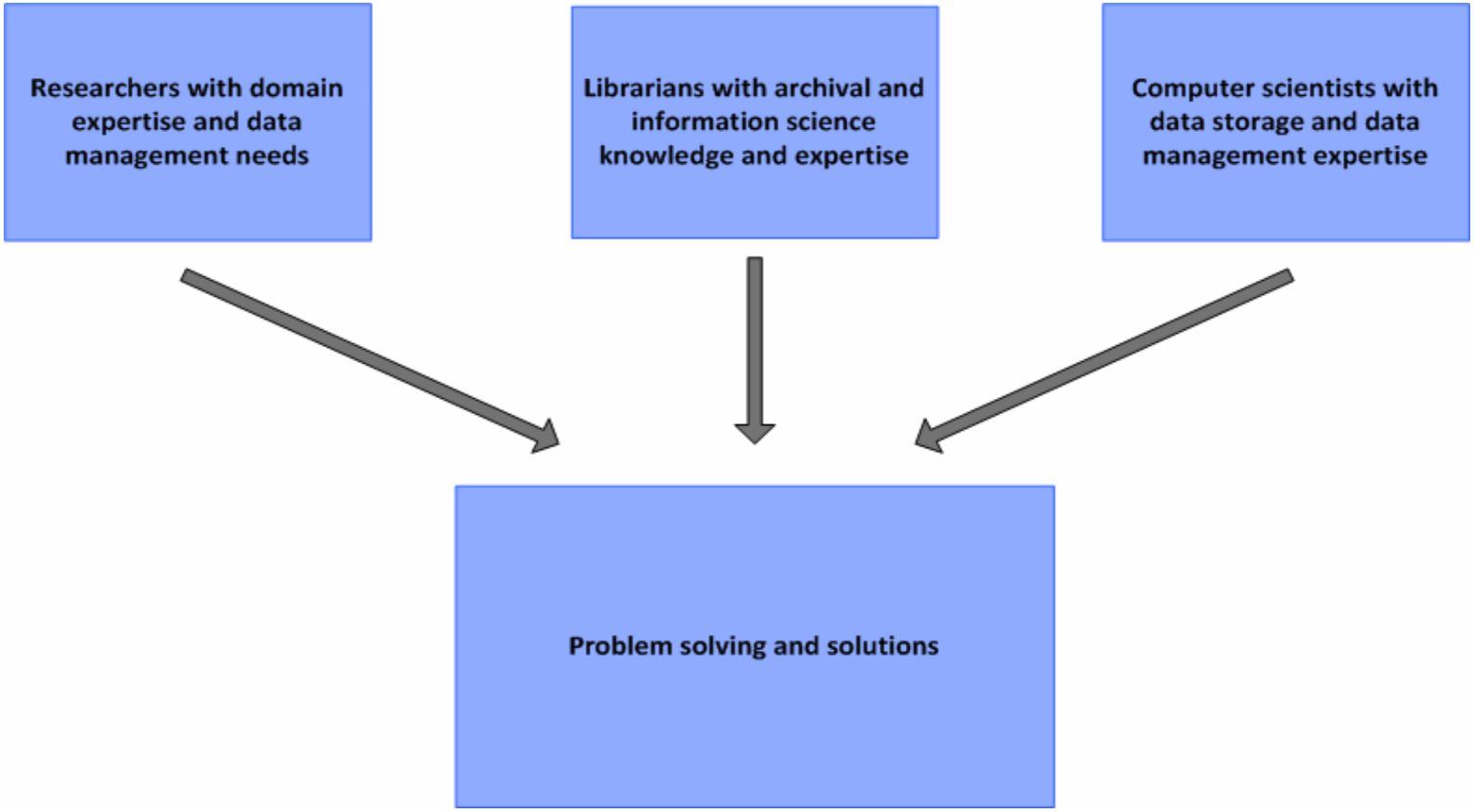
True Across the Disciplines

- **STEM Science**
  - OOI, Neon, NEES, GEOSS, PDB, BIRN, HIS
- **Social Science**
  - ICPSR datasets, local researcher datasets, surveys
  - Digital Video and Audio Interviews
- **Arts and Humanities**
  - Maurizio Seracini's x-ray collection of art masters, UCSD-TV videos
- **Cultural institutions**
  - Library mass digitization projects (Google, MS, OCA), Web archiving, Local digitization

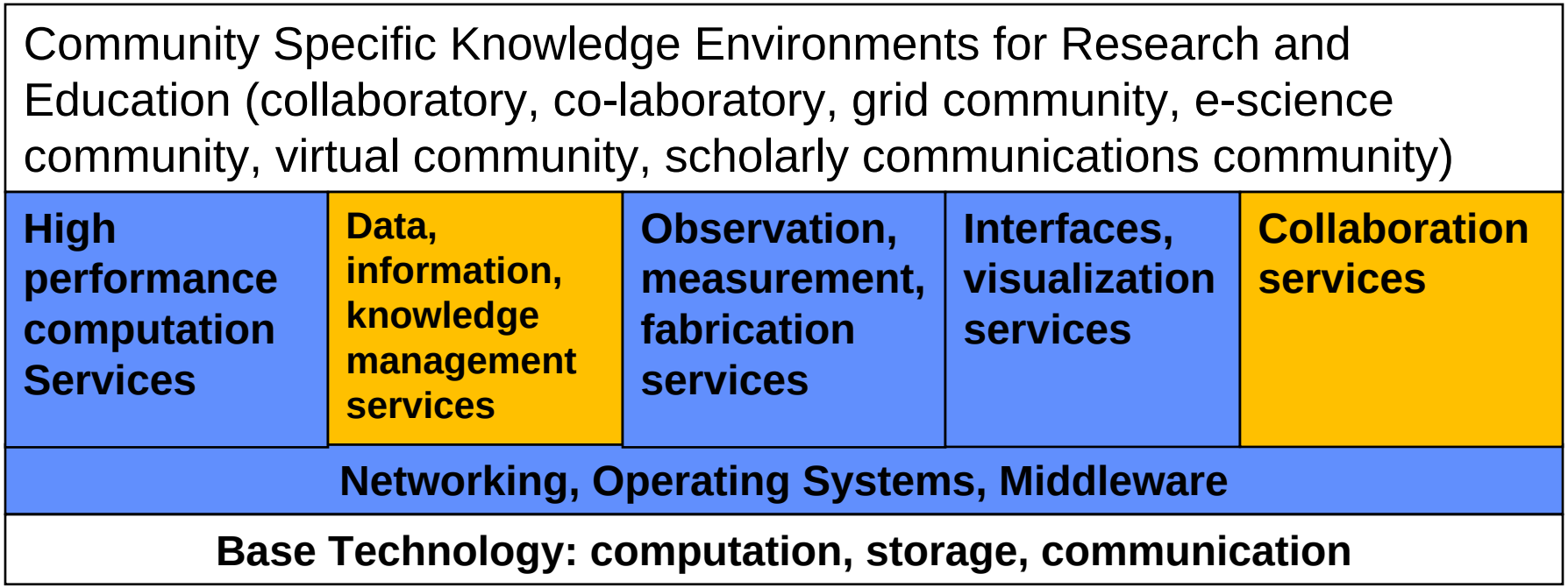
Urgent Need for:

- **Large-scale digital curation and preservation infrastructure**
- **Informed (metadata-/ontology-based) discovery of and access to data**
- **Links between the data and its research output**
- **Tools and services**
  - Data integration
  - Data mining
  - Data visualization

Urgent Need for: Collaboration



NSF (Atkins Report) CI Model



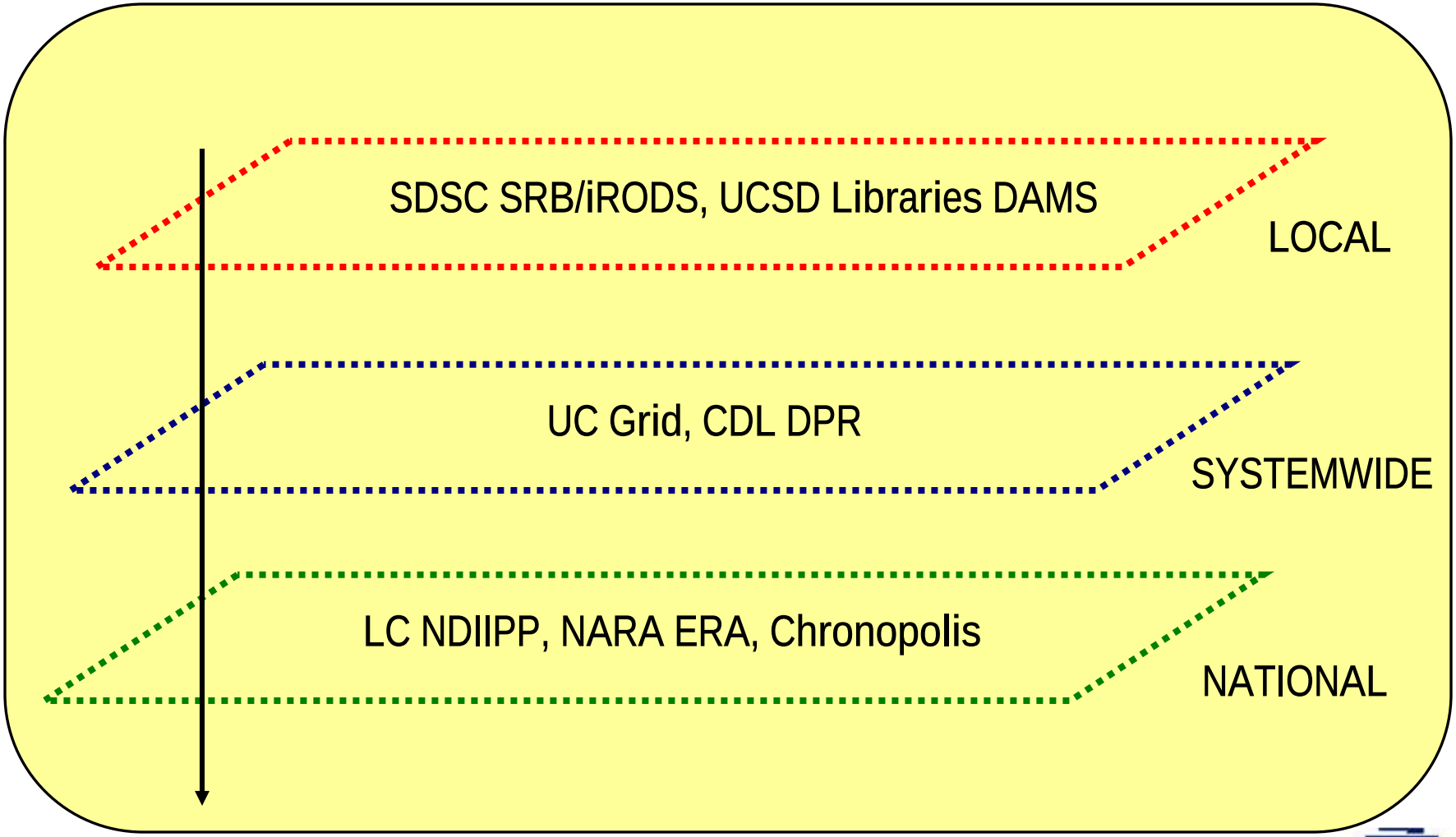
= Cyberinfrastructure: hardware, software, services, personnel, organizations

= Opportunities for Libraries

Collaborative Projects to Date

- **Collection Ingest**
  - **UCSD-TV videos, LC image collections, CDL Web archives**
- **Infrastructure**
  - **UCSD Libraries DAMS, CDL DPR, UC Grid, Chronopolis**
- **Interdisciplinary Data Integration**
  - **Neuroscience/Architecture databases**
- **Data Mining and Visualization**
  - **California Cooperative Oceanographic Fisheries Investigations (CalCOFI)**

UCSD Layers of CI Collaboration



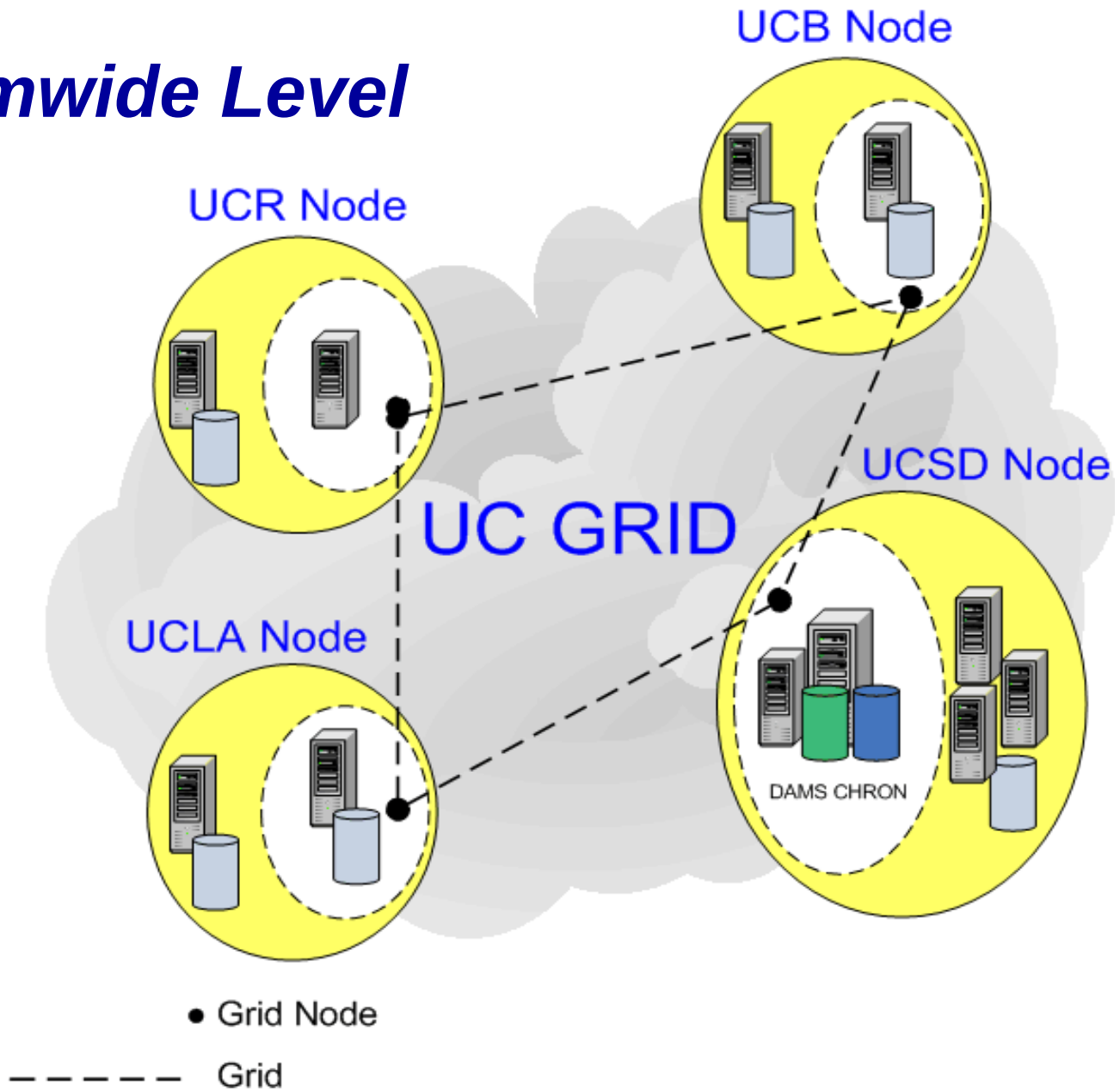
Local Level: DAMS and SDSC Storage

- **Collection identified**
- **Libraries Metadata Analysis and Specification Unit creates assembly plan**
  - Maps data to MODS, PREMIS, MIX, Local Schemas
- **Collection ingested with JETL (Java Extraction, Transformation, and Loading) Tool**
  - Original digital object
  - Assigned a unique, permanent identifier - ARK
  - Stored in SRB
  - Technical metadata extracted with JHOVE
  - Stored in SRB in under the same ARK
  - Metadata registered and stored in the SRB under the same ARK

DAMS Collections

- **Current Libraries Collections (6TB)**
  - Art Images
  - Electronic Theses and Dissertations
  - Special Collections objects:
    - Spanish Civil War Posters
    - *Dr. Seuss Went to War* Images
- **Future Data Collections**
  - Departmental Projects
  - Research Datasets
- **No Collection Too Big, No Collection Too Small**
  - RDF allows extensibility into any namespace

Systemwide Level



The UC Grid

- **Physical Connectivity**
  - 10 Gb among UC Campuses
- **UC Trust**
  - Shibboleth
  - Single Sign-On
- **Data Grid**
  - Google/OCA/Microsoft Books Project w/CDL
  - Mass Transit – data transfer between UC nodes
- **High Performance Computing**
  - Shared computing resources among UC campuses

National Level: Chronopolis Digital Preservation Program

- **Collaborative Initiative**
  - San Diego Supercomputer Center
  - University of California, San Diego Libraries
  - National Center for Atmospheric Research
  - University of Maryland, Inst. for Adv. Computer Studies
- **Long Term Digital Management and Preservation**
  - National center
  - Latest in storage technologies
  - Grid-enabled cyberinfrastructure
  - Operational data services
- **Research**

Chronopolis Collections

- **National Virtual Observatory (NVO)**
  - Currently 1 TB of Digital Palomar Observatory Sky Survey
- **Interuniversity Consortium for Political and Social Science Research (ICPSR)**
  - Currently 2 TB of Web-based data
  - Future plans include 10 TB of all ICPSR data collections
- **California Digital Library (CDL)**
  - Future Plans include 25 TB of Web-at-Risk crawl collection
- **Library of Congress (LC)**
  - Currently 2 TB of Prokudin-Gorskii image collection

Digital Preservation Data Grid

Administration for Policy and Outreach

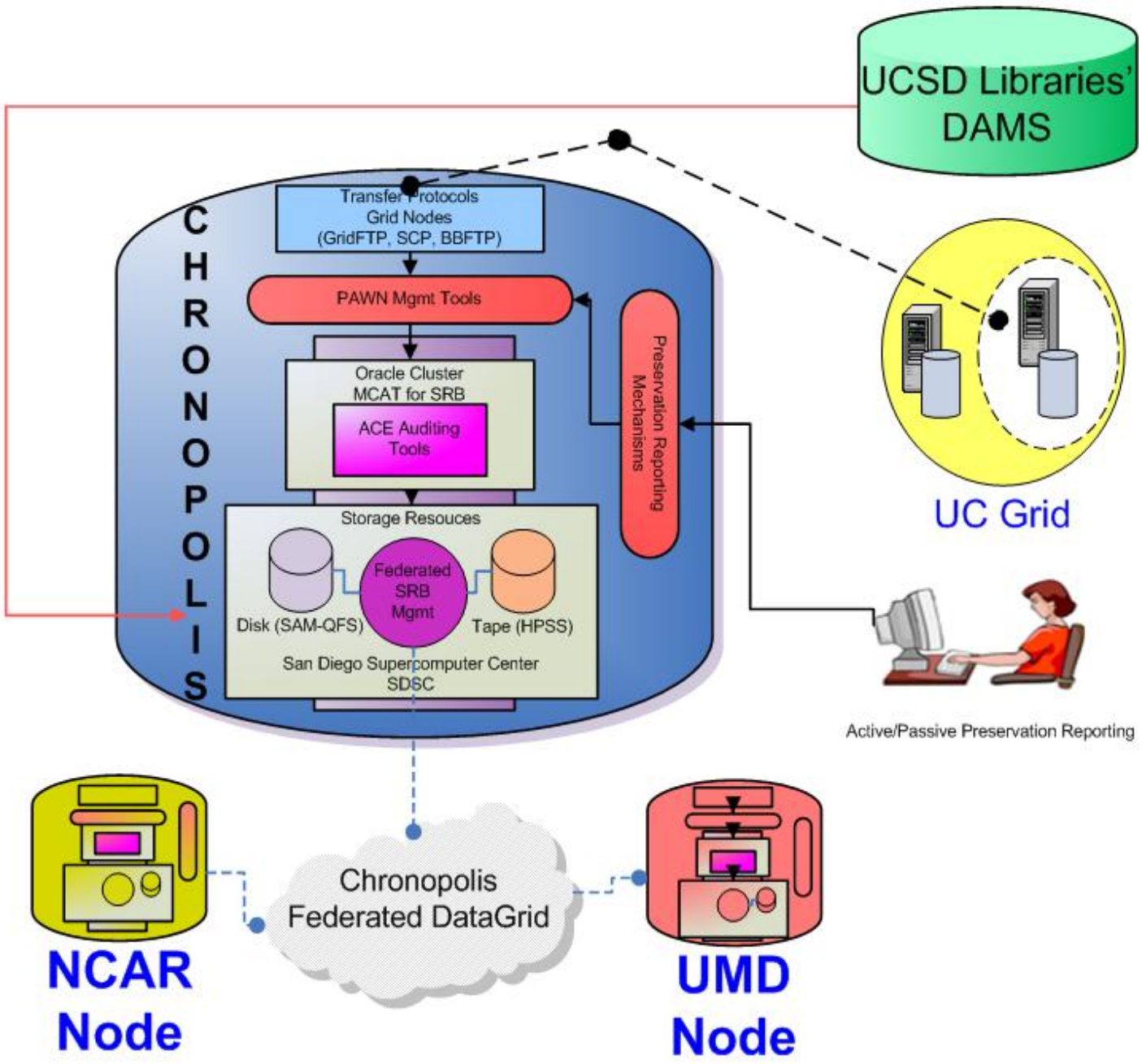
(Supports the overall partnerships and mgmt for preservation services and works as a liaison with Chroonpolis partners and other regional and national preservation programs)

Research and Development

(Research and development for rules-based preservation mgmt and technology forecasting for continual technology migration and mgmt)

Production Digital Preservation

(Long-term preservation with geographic replications and preservation services)



What Have We Learned?

- **We do indeed need each other**
- **Libraries bring a lot to the table**

What Libraries Can Contribute

- **Data acquisition, ingest layer**
  - Selection, taxonomy, ontology, metadata, workflow
- **Preservation layer**
  - Archival retention, format migration, QA, trust
- **Physical layer**
  - Storage, network, security, reliability standards
- **Service layer**
  - Discovery, retrieval, data mining, data visualization
- **Management layer**
  - Administration, budget, policy development

Competencies Leveraged

Faculty	Libraries	SDSC
❑ Domain expertise ❑ Data collection ❑ Taxonomies ❑ Ontologies ❑ Data mining ❑ Data reuse	❑ Archiving ❑ Metadata management ❑ Discovery-tool building ❑ Culture of service ❑ Culture of trust ❑ Project Management	❑ Grid storage ❑ Grid services ❑ Data management ❑ Data preservation ❑ Format migration

What Have We Learned?

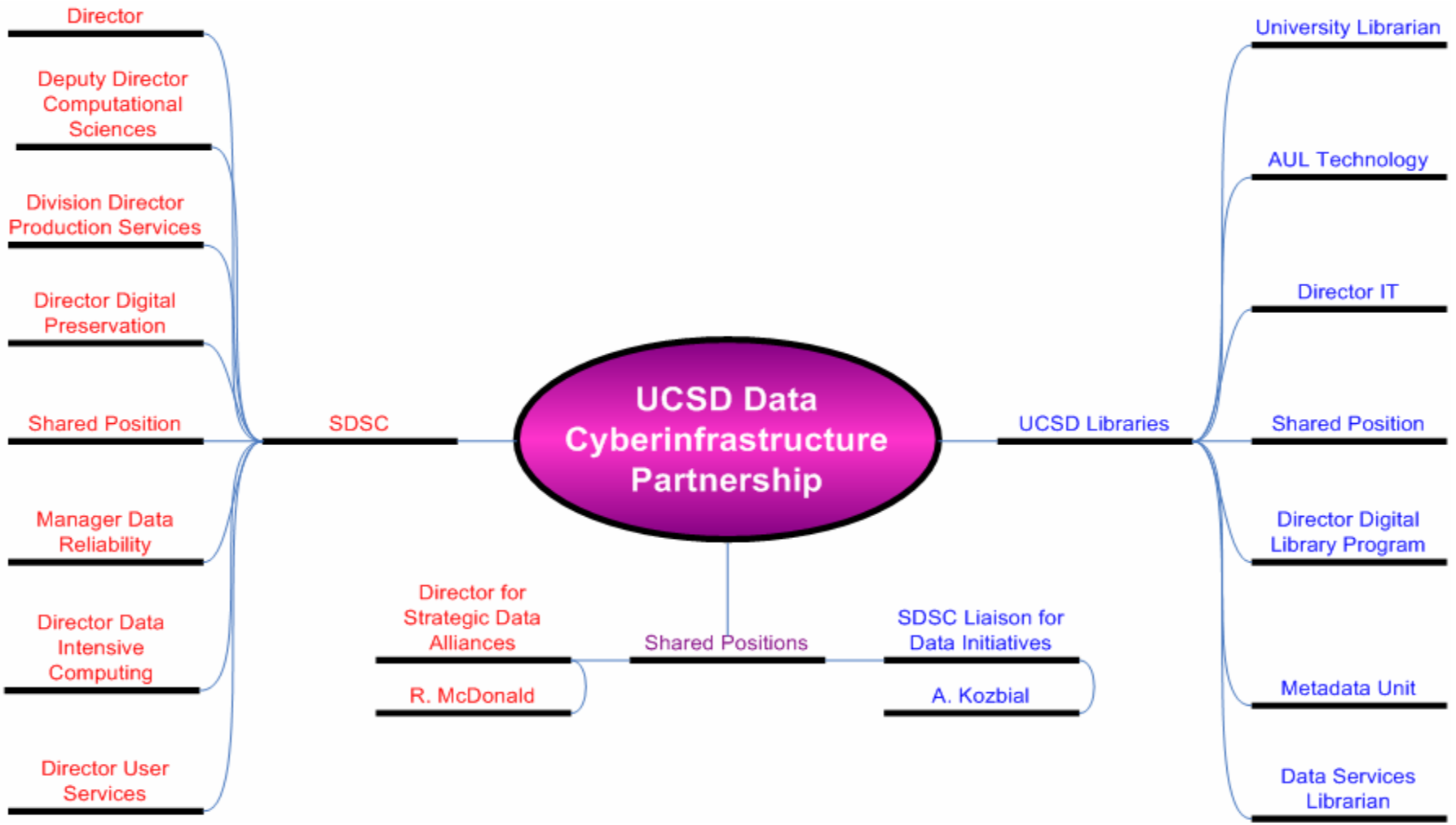
- **We do indeed need each other**
- **Libraries bring a lot to the table**
- **Substantial organizational differences**
- **New organizational structure would help**

Atkins Model for Collaborative DCI

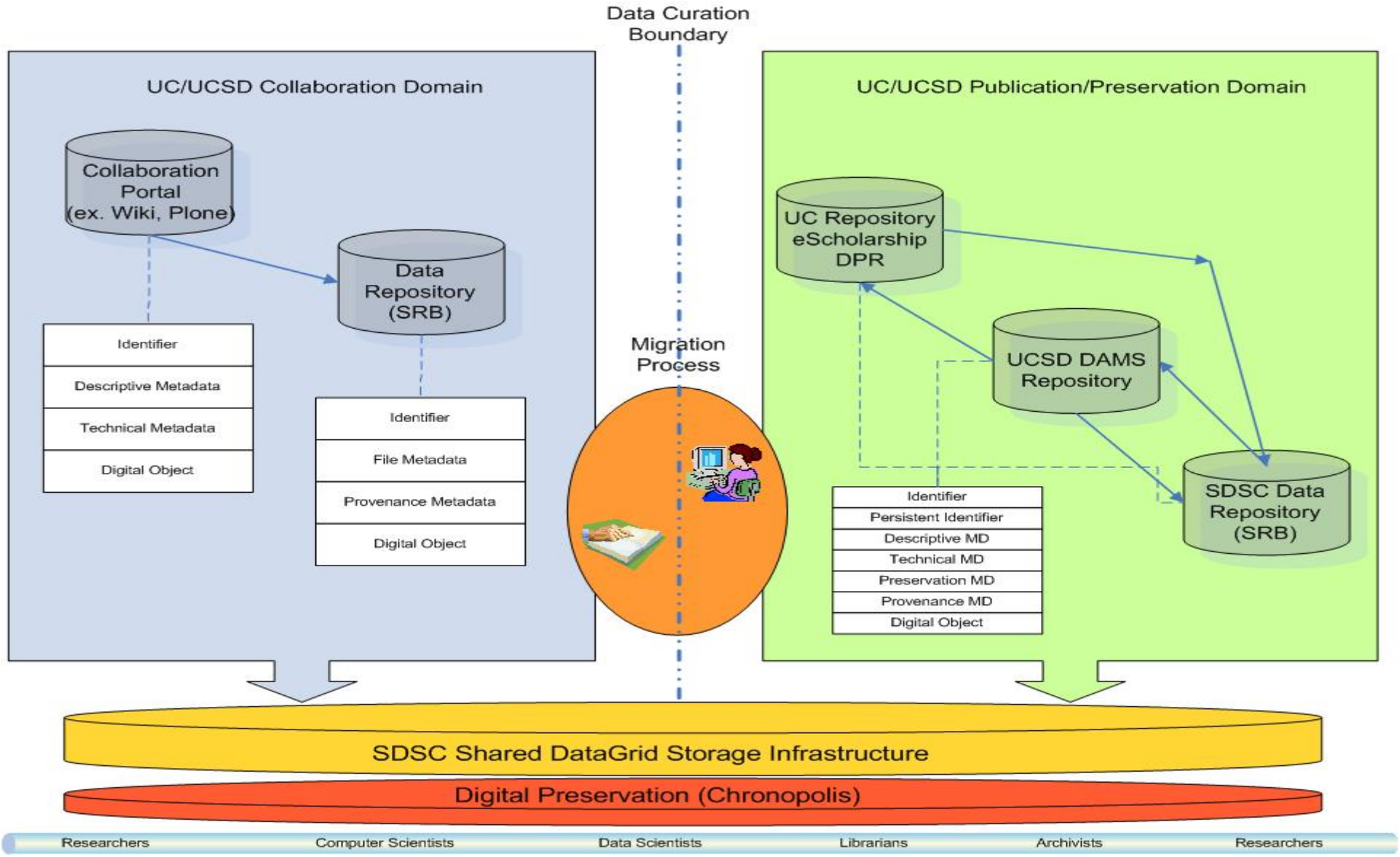
… new types or organizations … [that] … will integrate library and archival sciences, cyberinfrastructure, computer and information sciences, and domain science expertise …

— DataNet Program Solicitation NSF 07-601

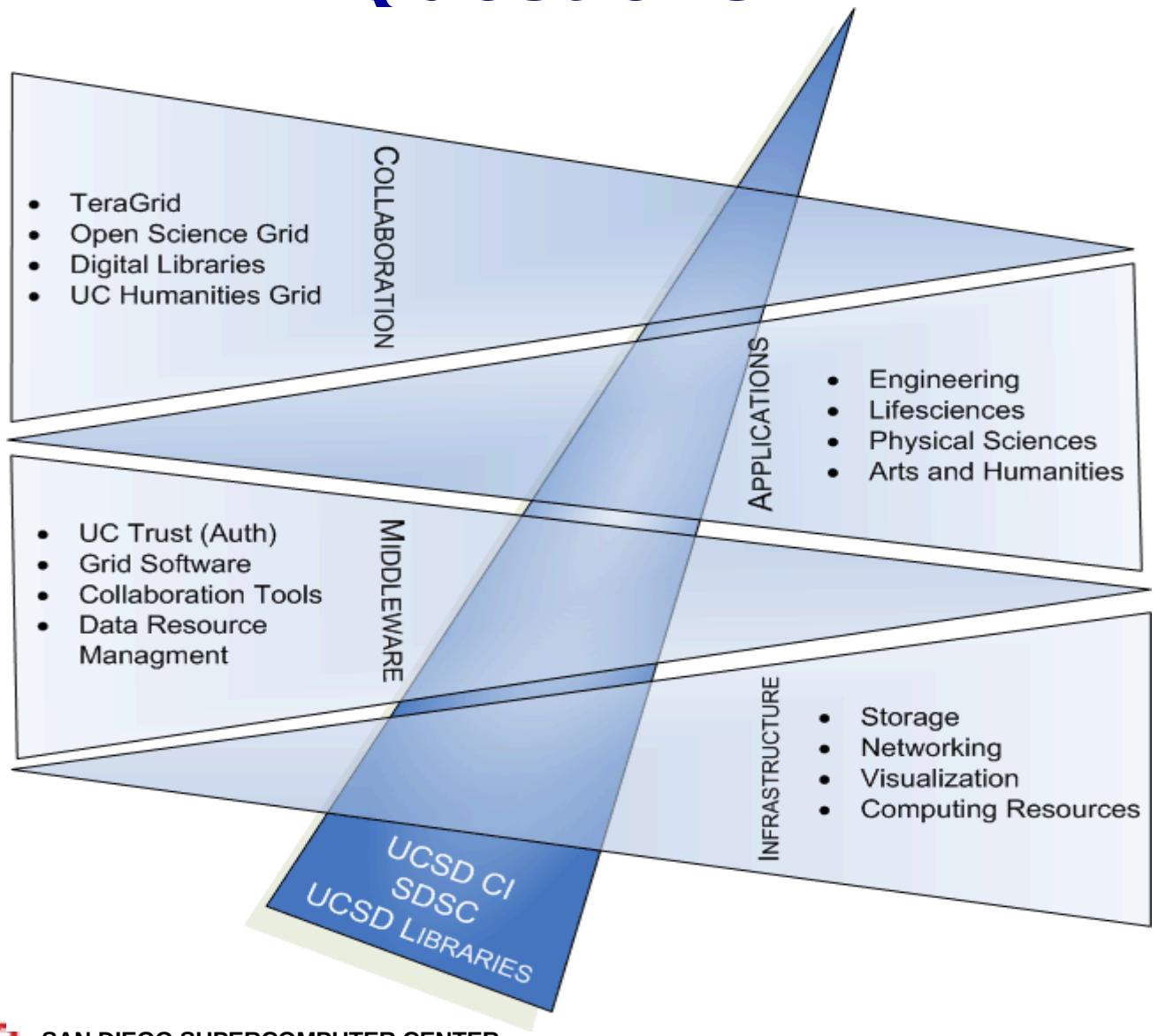
Collaboration at UCSD: Organization



Collaboration at UCSD: Program



Questions?


COLLABORATION
TeraGrid
Open Science Grid
Digital Libraries
UC Humanities Grid
APPLICATIONS
Engineering
Lifesciences
Physical Sciences
Arts and Humanities
MIDDLEWARE
UC Trust (Auth)
Grid Software
Collaboration Tools
Data Resource Managment
INFRASTRUCTURE
Storage
Networking
Visualization
Computing Resources
UCSD CI
SDSC
UCSD LIBRARIES