Survey of Campus Research Storage Needs, etc.

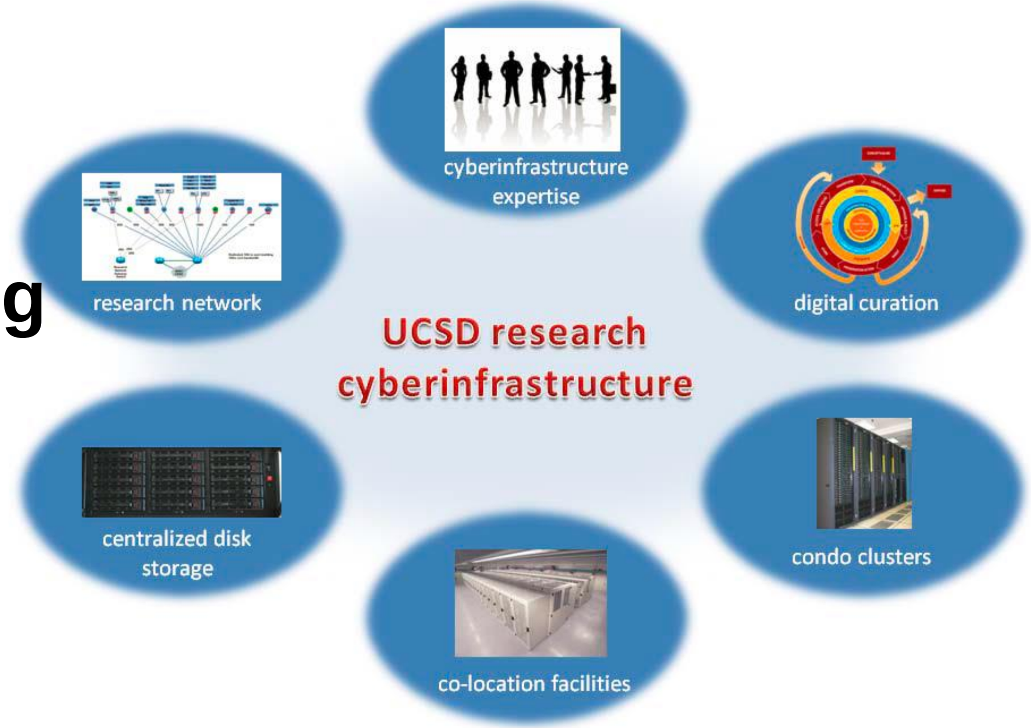
Sustainability Workshop

October 2, 2013

Richard Moore

Elements of UCSD's Integrated Research CyberInfrastructure (RCI) Program

- Data Center Colocation
- Networking
- Research Computing
- Centralized Storage
- Data Curation
- Technical Expertise



rci.ucsd.edu

Campus Survey of Researchers' Data Requirements

- Conducted survey of a broad sample of ~50 representative PIs to understand technical and cost requirements
- An additional motive was to increase awareness of the RCI program

- Identify common needs, and define sustainable RCI business model with strong adoption
- Develop centralized, production storage services

PI Interview Responses: Where is Your Data Coming From?

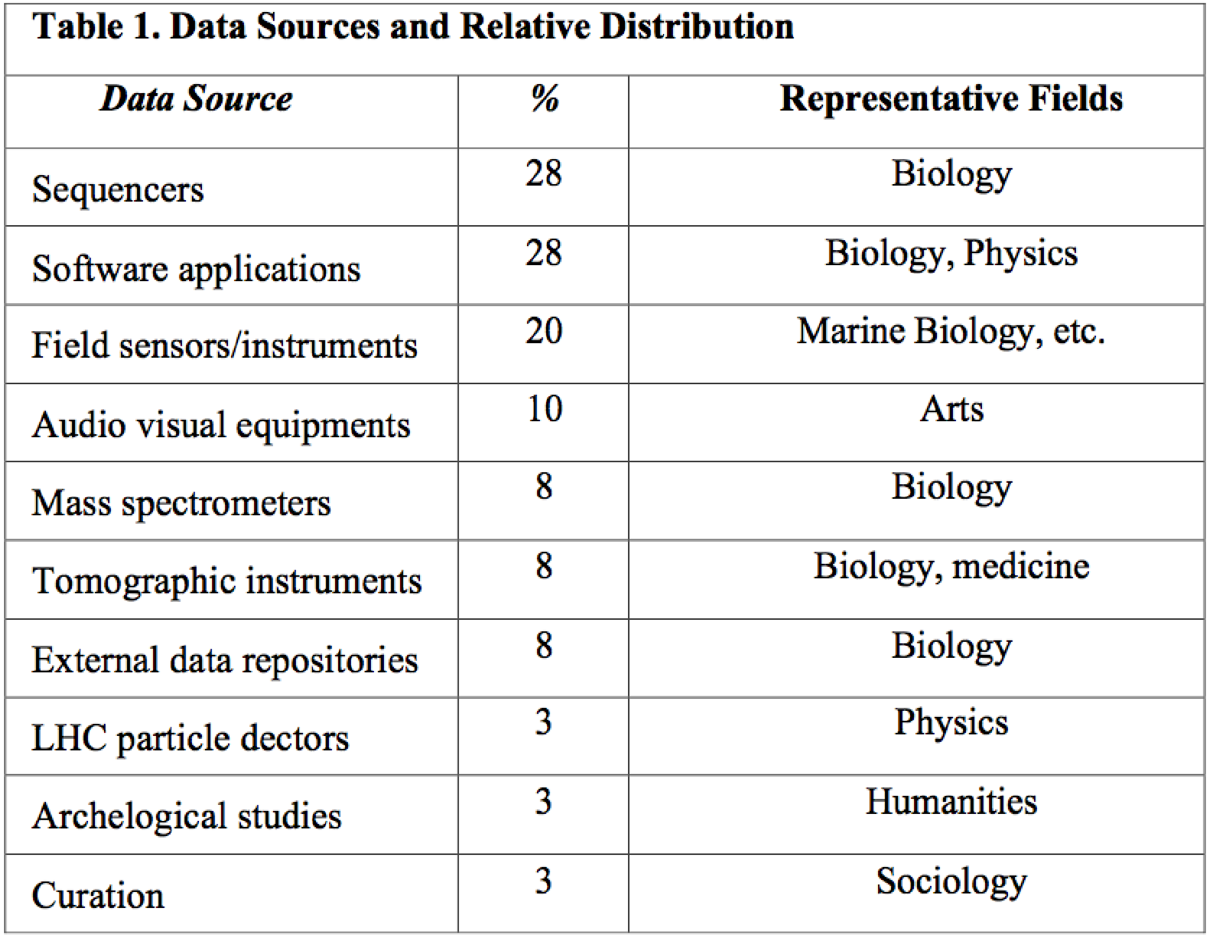


Table 1. Data Sources and Relative Distribution

Data Source	***%***	**Representative Fields**
Sequencers	28	Biology
Software applications	28	Biology, Physics
Field sensors/instruments	20	Marine Biology, etc.
Audio visual equipments	10	Arts
Mass spectrometers	8	Biology
Tomographic instruments	8	Biology, medicine
External data repositories	8	Biology
LHC particle dectors	3	Physics
Archelogical studies	3	Humanities
Curation	3	Sociology

Numbers reflect percentages of PIs surveyed that utilize each solution ;
Individual PIs use multiple solutions, so %'s add up to >100%.

- **Indicates use cases for storage and connectivity requirements**
- **Data sources:**
  - ~50% campus instruments
  - ~30% simulations (XSEDE, campus, lab systems)
  - ~20% field instruments
  - ~15% other external sources
- **%'s reflect PIs, not data volume**

How do You Handle Data Storage/Backup?

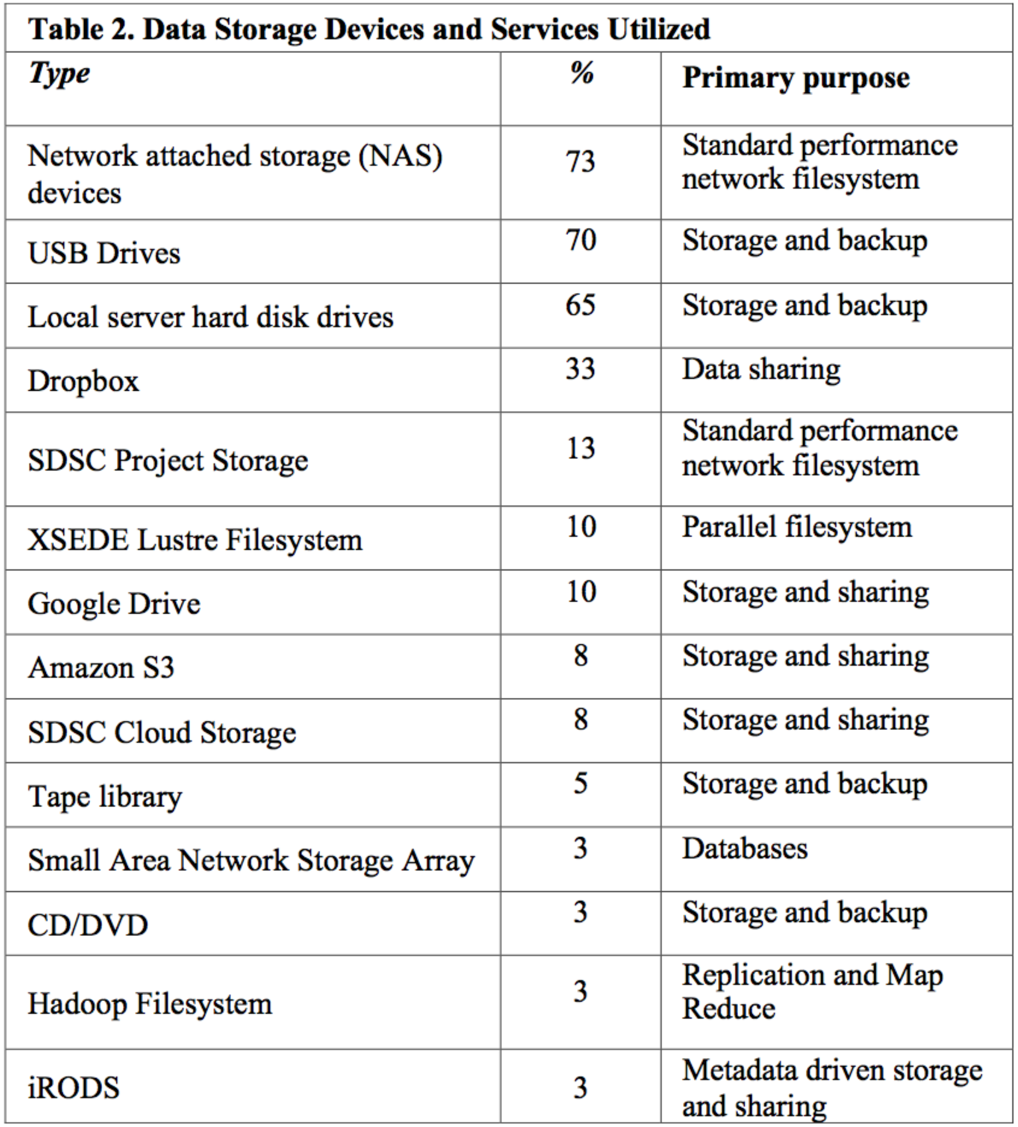


Table 2. Data Storage Devices and Services Utilized

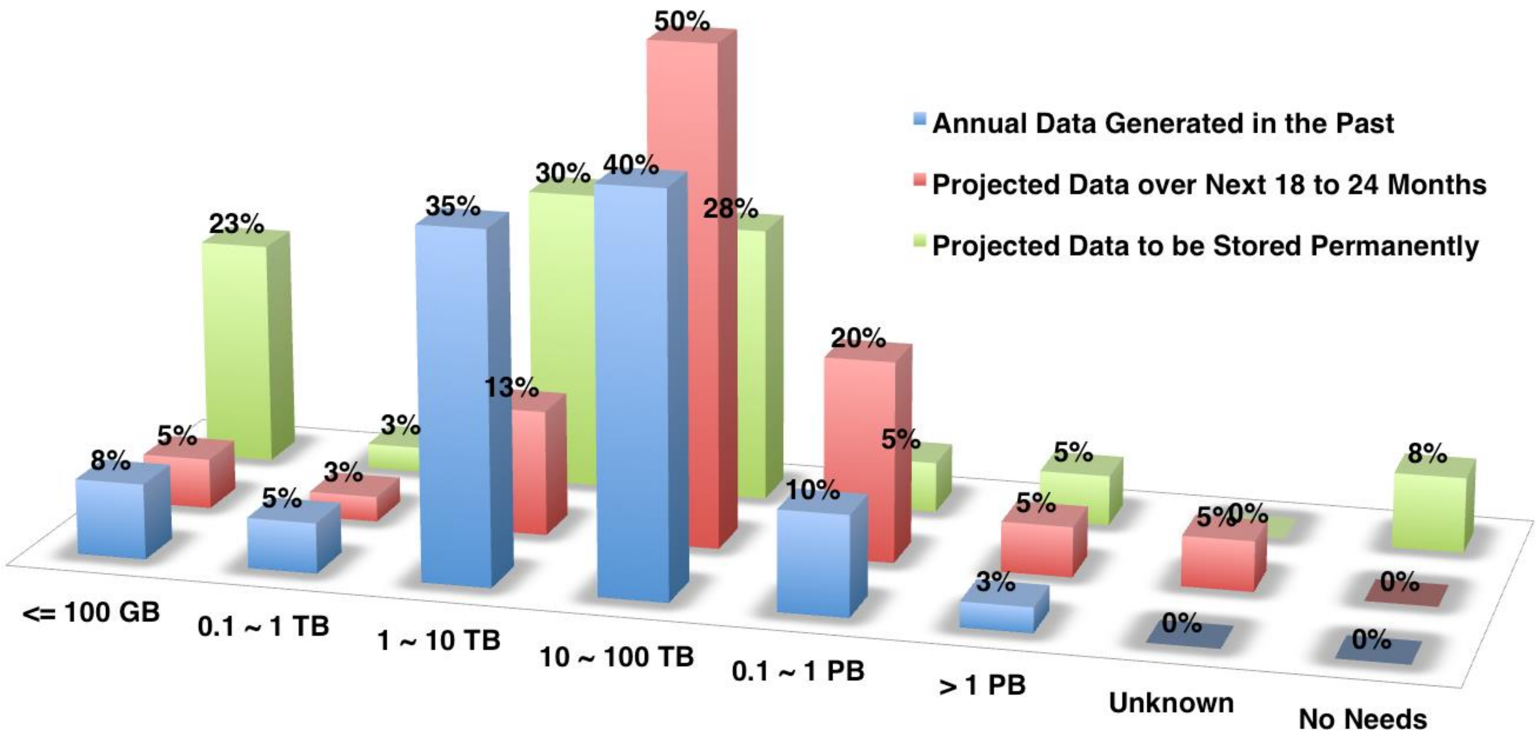
Type	*%*	**Primary purpose**
Network attached storage (NAS) devices	73	Standard performance network filesystem
USB Drives	70	Storage and backup
Local server hard disk drives	65	Storage and backup
Dropbox	33	Data sharing
SDSC Project Storage	13	Standard performance network filesystem
XSEDE Lustre Filesystem	10	Parallel filesystem
Google Drive	10	Storage and sharing
Amazon S3	8	Storage and sharing
SDSC Cloud Storage	8	Storage and sharing
Tape library	5	Storage and backup
Small Area Network Storage Array	3	Databases
CD/DVD	3	Storage and backup
Hadoop Filesystem	3	Replication and Map Reduce
iRODS	3	Metadata driven storage and sharing

Numbers reflect percentages of PIs surveyed that utilize each solution ; Individual PIs use multiple solutions, so %'s add up to >100%.

- **Storage Devices**
  - Network accessible storage (NAS), USB and server local drives dominate
  - Use of Dropbox for sharing
  - Others use Google Drive, Hadoop, XSEDE, SDSC co-location
- **Backup modes**
  - Replicated copies in two NAS
  - A copy in the NAS,
  - A copy in local hard drive (laptop/workstation),
  - And a copy in a USB drive
  - Maybe a copy in email/Dropbox
- **Problems:**
  - Out of sync
  - Lost track of its location
  - Lost version control
  - High cost of recovery

How much storage do you need: now, future, permanently?

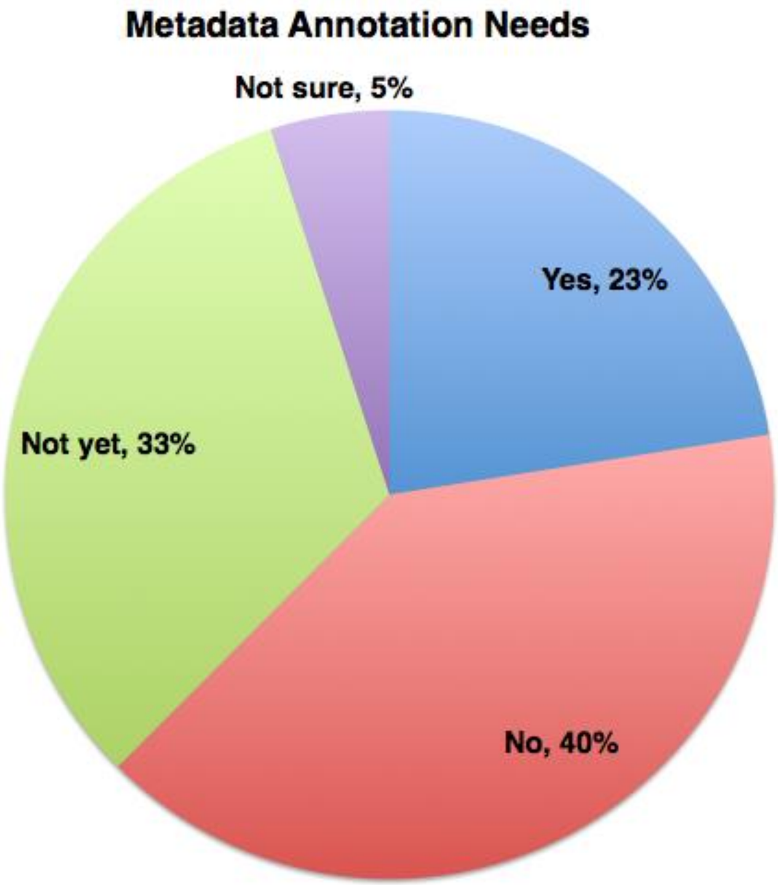
Data Storage and Growth in the Present and Next 2 Years

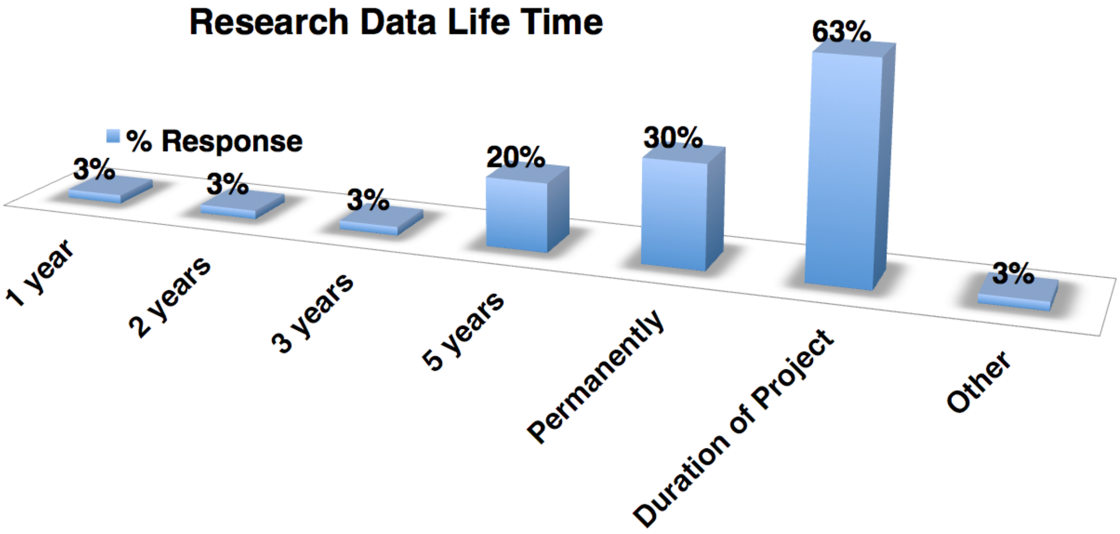


- **For PIs interviewed, current needs 1-1000TB**
- **Increasing in future**
- **Perceptions of permanent storage interesting – none for some, intermediate for many, large for a few**

Metadata and retention requirements

Do you need metadata annotation capability?





How long do you need to retain your data?

Top Requirements for Campus Cyberinfrastructure

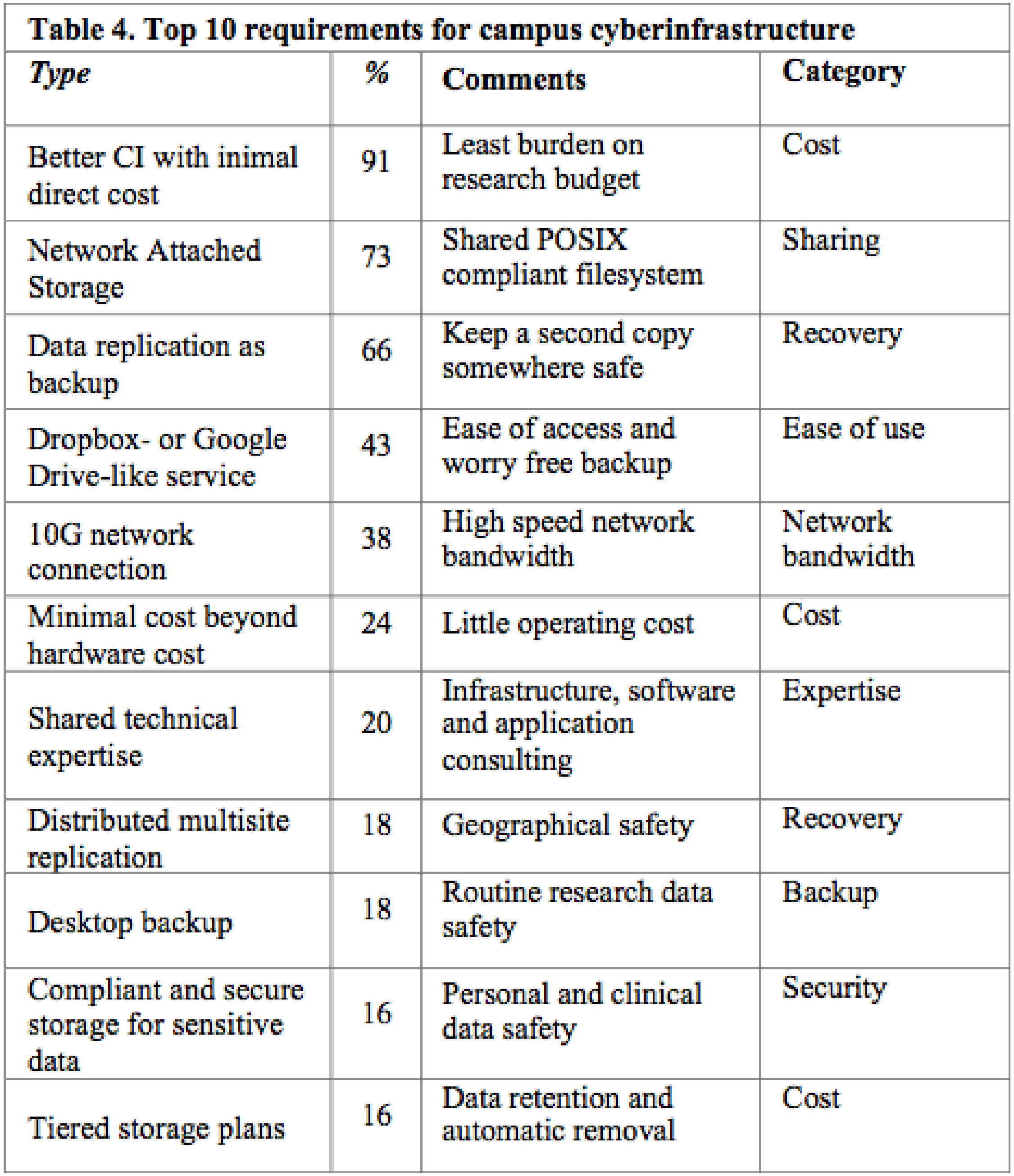


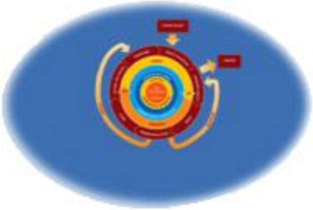
Table 4. Top 10 requirements for campus cyberinfrastructure

Type	%	Comments	Category
Better CI with inimal direct cost	91	Least burden on research budget	Cost
Network Attached Storage	73	Shared POSIX compliant filesystem	Sharing
Data replication as backup	66	Keep a second copy somewhere safe	Recovery
Dropbox- or Google Drive-like service	43	Ease of access and worry free backup	Ease of use
10G network connection	38	High speed network bandwidth	Network bandwidth
Minimal cost beyond hardware cost	24	Little operating cost	Cost
Shared technical expertise	20	Infrastructure, software and application consulting	Expertise
Distributed multisite replication	18	Geographical safety	Recovery
Desktop backup	18	Routine research data safety	Backup
Compliant and secure storage for sensitive data	16	Personal and clinical data safety	Security
Tiered storage plans	16	Data retention and automatic removal	Cost

- **Cost effectiveness tops list**
- **Ease of use follows**
- **“Cost is King, Ease of Use Follows”**
- **Reliable, NFS/CIFS storage most common platform**
- **Many responses relate to data durability – backups/ copies/tiered storage**
- **High-speed networking enhances quality of service**
- **“Compliant” environment (storage/computing)**
- **Tiered storage options is desirable**

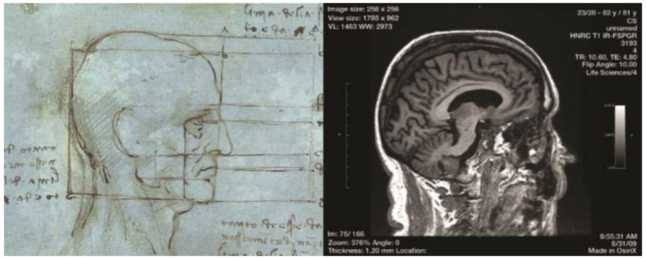
Research Computing (in production now)

- **RCI is evolving SDSC's Triton system to the "*Triton Shared Computing Cluster*" (TSCC)**
- **Condo model: Researchers purchase compute nodes which are operated as part of shared cluster for 3-4 years**
  - PI buys hardware & modest ops fee
  - Lower ops cost than local PI cluster; larger-scale resource; professionally-managed
- **Hotel: Purchase time by the core-hour; shared queue**



Data Curation – in pilot (production FY13-14)

- **Completing a two-year pilot phase**
  - How do lab personnel work with librarians to curate their data?
  - How much work is required to curate data and what are options?
  - What is a sustainable business model for curation within RCI project?
- **Five representative programs across UCSD selected as pilots**
  - The Brain Observatory (Annese)
  - Open Topography (Baru)
  - Levantine Archaeology Laboratory (Levy)
  - SIO Geological Collections (Norris)
  - Laboratory for Computational Astrophysics (Wagner)



- **Using existing tools whenever possible**
  - Storage at SDSC, campus high-speed networking, Digital Asset Management System (DAMS) at UCSD Libraries, Chronopolis digital preservation network
- **Also, develop Data Management Plan tools and provide training**
- **Anticipate production curation services in FY13-14**

Some Comments and Lessons Learned

- **Campus multi-year budget commitments make a difference to adoption – obvious but …**
- **In-person interactions very important to adoption**
- **Wish we'd hired an expert in conducting survey**
- **Comment yesterday re campus requiring that PIs put skin in the game – not only $, but litmus test**
  - However, makes it hard to plan and prepare for 3-5 years out
- **'Economies of scale' leverage varies for different services (e.g. colocation -> data curation)**
- **UC systemwide pilot project (may also apply to some regional collaborations) - getting one person to say yes is a lot easier than N people**