UT Southwestern Medical Center
Lyda Hill Department of Bioinformatics

BioHPC

Lab Data Management

A user perspective

March 22, 2023

Agenda

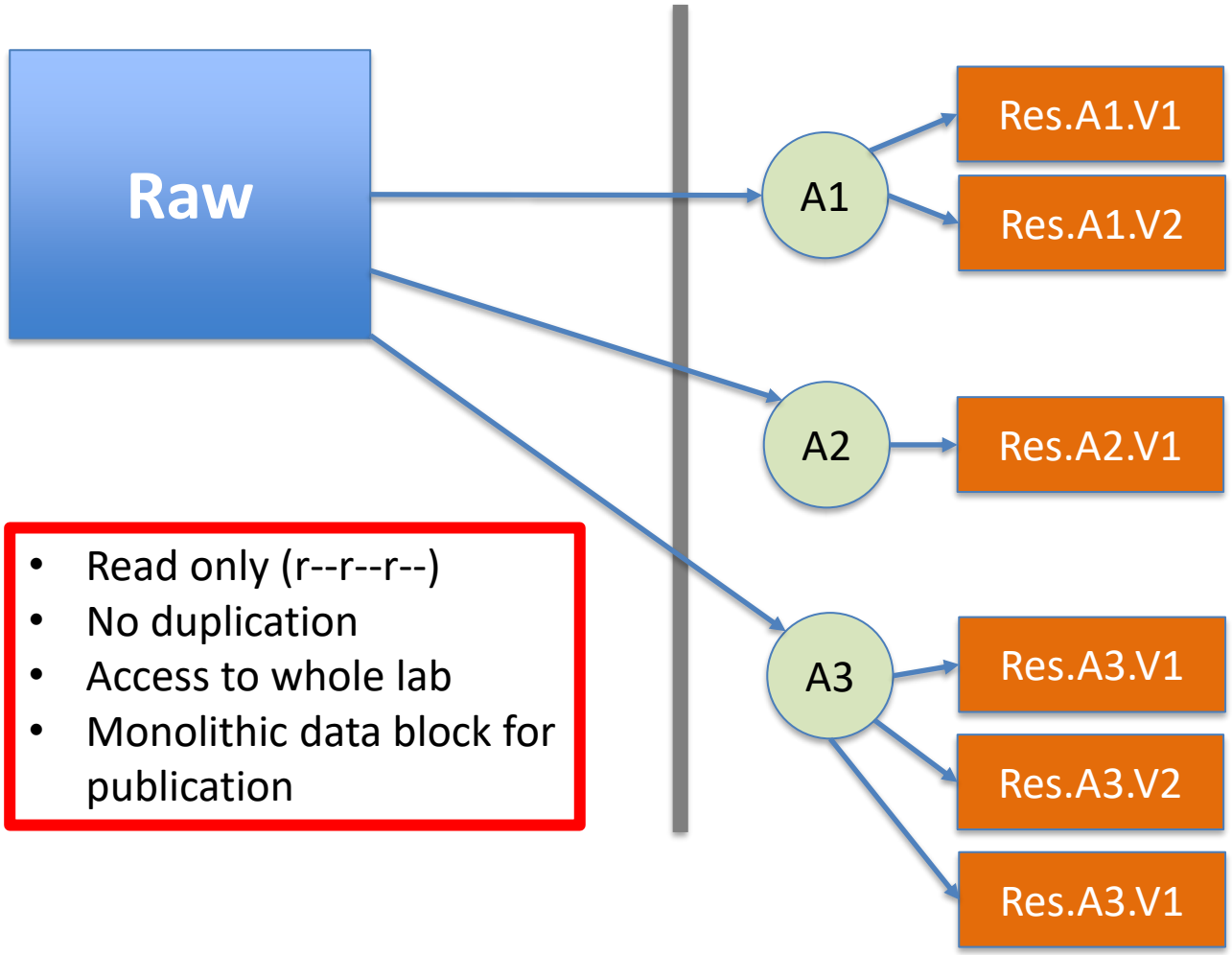
- Challenges in implementing robust data management practices – what to look for
- Ideas on implementing such practices on BioHPC
- Interfacing lab data management strategies with data sharing requirements by journals and funders
- A word or two on NIH's data management and sharing requirements
- A word on data retention
- The great new promise: BioHPC's plans for a 3-tier storage system

Danuser/Fiolka labs use case

- 500 – 700 TB of image data
  - ➢ Large files and large stacks of small files
- 30 users
- Mix of commercial software, open-source packages, home-written software
  - ➢ Communication between packages via filesystem
- Multiple separable projects
- Each project involves
  - ➢ 1 or several data generators
  - ➢ 1 or several data analysts with *distinct* research questions
- Same data shared over multiple publications
- Data reuse over generations of lab members / trainees
- Large intermediate result files
  - ➢ Quasi-duplication of data

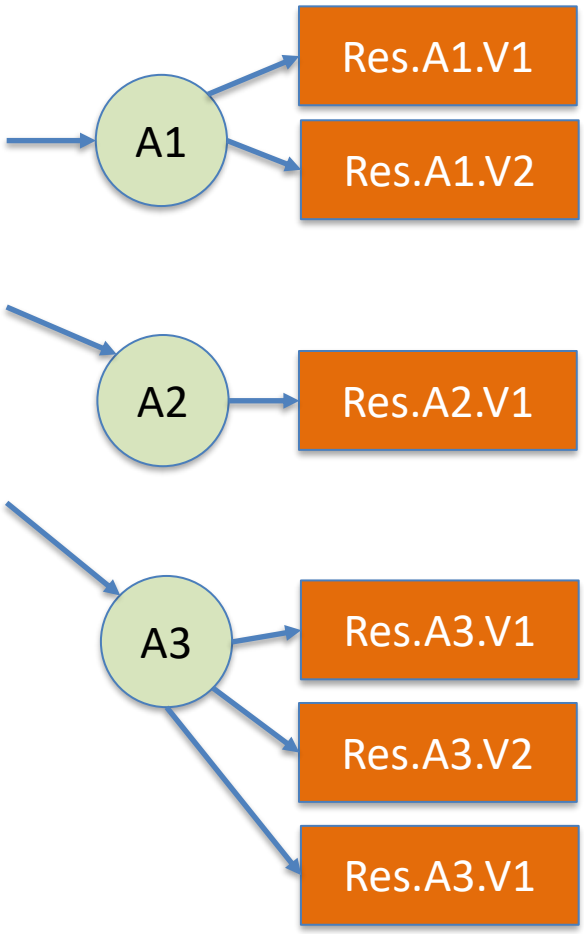
3 Rules for Lab Data Management

Rule #1 – Separation of raw and processed data



- Read only (r--r--r--)
- No duplication
- Access to whole lab
- Monolithic data block for publication

Rule #1 – Separation of raw and processed data



- Set to rw-r--r--
- Each lab member controls personal results
- [Results can be shared between lab members as sub blocks]
- Obsolete processing trees and/or final results can be deleted
- Processing trees/Results of departed lab members can be integrally deleted without affecting raw data or still active processing trees by other users

Rule #1 – Implementation on File System

Scenario with no result sharing between lab members

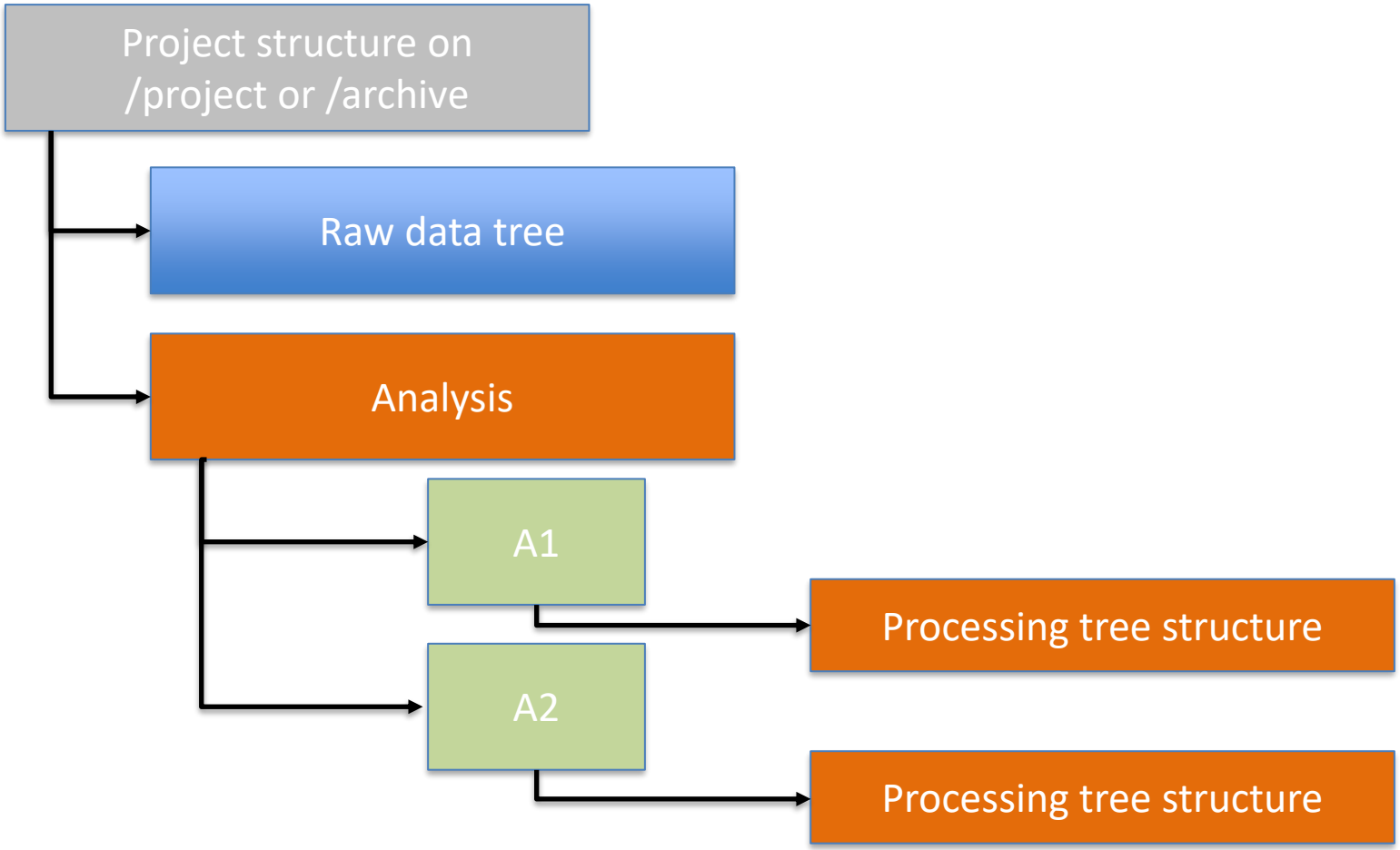
Project structure on /project or /archive

Processing tree structure on /work

!!!! Deleted with user departure

Rule #1 – Implementation on File System

Scenario *with* result sharing between lab members and automatic result longevity



Rule #2 – Separation of projects

Project structure on /project or /archive

- Monolithic blocks
  - Move
  - Zip
  - Delete
  - Symbolic linking
- If needed, lab sub-groups
- Separate documentation
- Usage monitoring
  - Data cleaning

```
2.0T    /project/bioinformatics/Danuser_lab/mechanometabolism
1.5T    /project/bioinformatics/Danuser_lab/P01biosensor
137G    /project/bioinformatics/Danuser_lab/MultispectralMicroscope
3.9T    /project/bioinformatics/Danuser_lab/3DTPE
9.0T    /project/bioinformatics/Danuser_lab/shared
14T     /project/bioinformatics/Danuser_lab/P01if
151G    /project/bioinformatics/Danuser_lab/danuser_ci
8.6T    /project/bioinformatics/Danuser_lab/P01adhesion
59T     /project/bioinformatics/Danuser_lab/3Dmorphogenesis
8.7T    /project/bioinformatics/Danuser_lab/ActinGrangerCausality
8.0K    /project/bioinformatics/Danuser_lab/softwareDevelopment
45T     /project/bioinformatics/Danuser_lab/microscopeDevelopment

64T     /archive/bioinformatics/Danuser_lab/zebrafish
3.5T    /archive/bioinformatics/Danuser_lab/liveCellHistology
2.4T    /archive/bioinformatics/Danuser_lab/lungCancer
125T    /archive/bioinformatics/Danuser_lab/melanoma
45T     /archive/bioinformatics/Danuser_lab/liveCellHistology_project
1.6T    /archive/bioinformatics/Danuser_lab/microscopeDevelopment
20T     /archive/bioinformatics/Danuser_lab/Ras
8.0K    /archive/bioinformatics/Danuser_lab/mechanometabolism
8.7T    /archive/bioinformatics/Danuser_lab/GEFscreen
3.2T    /archive/bioinformatics/Danuser_lab/softwareDevelopment
106G    /archive/bioinformatics/Danuser_lab/publications
7.6T    /archive/bioinformatics/Danuser_lab/3Dmorphogenesis
1.7T    /archive/bioinformatics/Danuser_lab/shared
1.4T    /archive/bioinformatics/Danuser_lab/externBetzig
332T    /archive/bioinformatics/Danuser_lab/Fiolka
```

BioHPC assistance required

Rule #3 – Data and result documentation

Raw

1. Maximum: Database for management, e.g. OMERO, …
   - BUT, ensure direct access via filesystem
2. Use file formats with embedded metadata
   - Choose non-proprietary formats
   - Maintain separate data content table, e.g. in an electronic lab notebook (like LabArchives)
3. Maintain separate data content table with metadata information
   - Risk: Inconsistencies due to manual documentation
4. Minimum: File naming convention
   - Encode key meta data in filename
   - Document convention in separate location like LabArchives
   - Document metadata in Readme files attached to each file tree leaflet

Documentation must be intelligible for multiple generations of lab members

Rule #3 – Data and result documentation

Analysis

1. Maximum: Use software with integrated workflow management
2. Minimum: Maintain log-files documenting every call
   - You are responsible for the reconstruction of the entire analytical path from raw data to result
   - Software version control and containerization

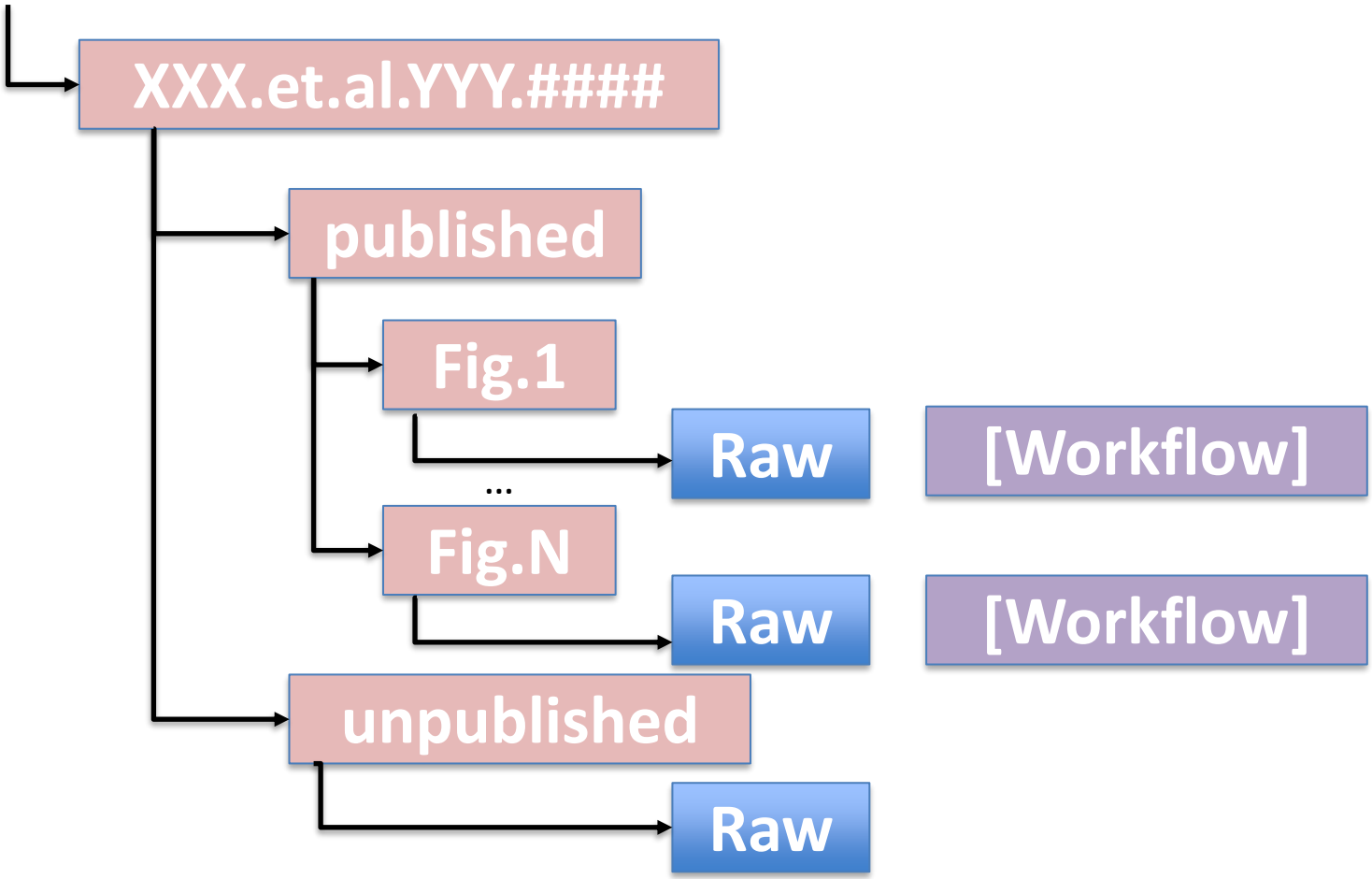
Meeting publisher's increasingly stricter data sharing requirements

Mandate(s):

- Any raw data feeding into a result figure must be accessible through a repository with a doi
- Standard result data types must be deposited in specialized archives
- [Processing workflows must be documented, with a doi]
- Home-grown software packages must be accessible via Github or even as frozen version with a doi

Tactics for organizing published data

/archive/bioinformatics/Danuser_lab/publications



Tactics for organizing published data

1. Deposit /published in a public repository for unstructured data
   - Zenodo, Mendeley, ..., Texas Data Repository
   - Fetch doi for paper
   - Deposited data constitutes long-term back-up of the high-value portion
2. Remove raw data copied into /published and /unpublished from project structure
   - Avoid duplication of old data sets
   - Publication is an implicit documentation of data to next generation lab members

Useful links to the Texas Data Repository

https://dataverse.tdl.org/dataverse/utswmed

https://utsouthwestern.libguides.com/utswrdr

https://utsouthwestern.libguides.com/utswrdr/quick-start-guide

Meeting NIH data sharing requirements

Next 3 slides courtesy of Dr. Joan Conaway

NIH Data Management and Sharing Policy

- What's required?
  - Must submit Data Management and Sharing Plan with application and have it approved by NIH staff.
  - Must comply with Plan.
- What data needs to be shared?
  - All data "commonly accepted as being of sufficient quality to validate and replicate findings."
  - Includes negative results.
- What doesn't need to be shared?
  - Lab notebooks, preliminary data, irreproducible or uninterpretable results, assay optimization…

https://sharing.nih.gov/faqs#/data-management-and-sharing-policy.htm

NIH Data Management and Sharing Policy

- Where does it need to be shared?
  - Ideally: Repository that is searchable, sustainable, has DOIs or accession numbers, supports metadata, free and easy access, allows re-use and citation of data.

- When does it have to be shared?
  - At publication or end of project period (grant close-out).
  - *Successful competitive renewal can extend project period.*

NIH Data Management and Sharing Policy

- Who is responsible for ensuring compliance with Data Management Plan?
  - Investigator but…
  - Institution is ultimately responsible.
  - Noncompliance may be factored into future funding decisions -

    not just for the investigator but for the institution.

Meeting NIH data sharing requirements

My interpretation:

- Meeting publisher mandates will automatically meet NIH requirements, with 2 exceptions:
  - Negative results -> dump /unpublished in a separate repository
  - Really unpublished results with a grant ending -> who cares?

Rolling out BioHPC's data storage v2.0

Reminder – Status Quo

User-centric
(will be removed with user departure)

Lab-centric
(long-term storage for research labs)

/home2
50G limit
Backup 2x/week

/work
50T limit
Backup 1x/week

/project
5T limit per PI as default **$$$**
Increase per PI request and chair approval
No backup (can be requested)

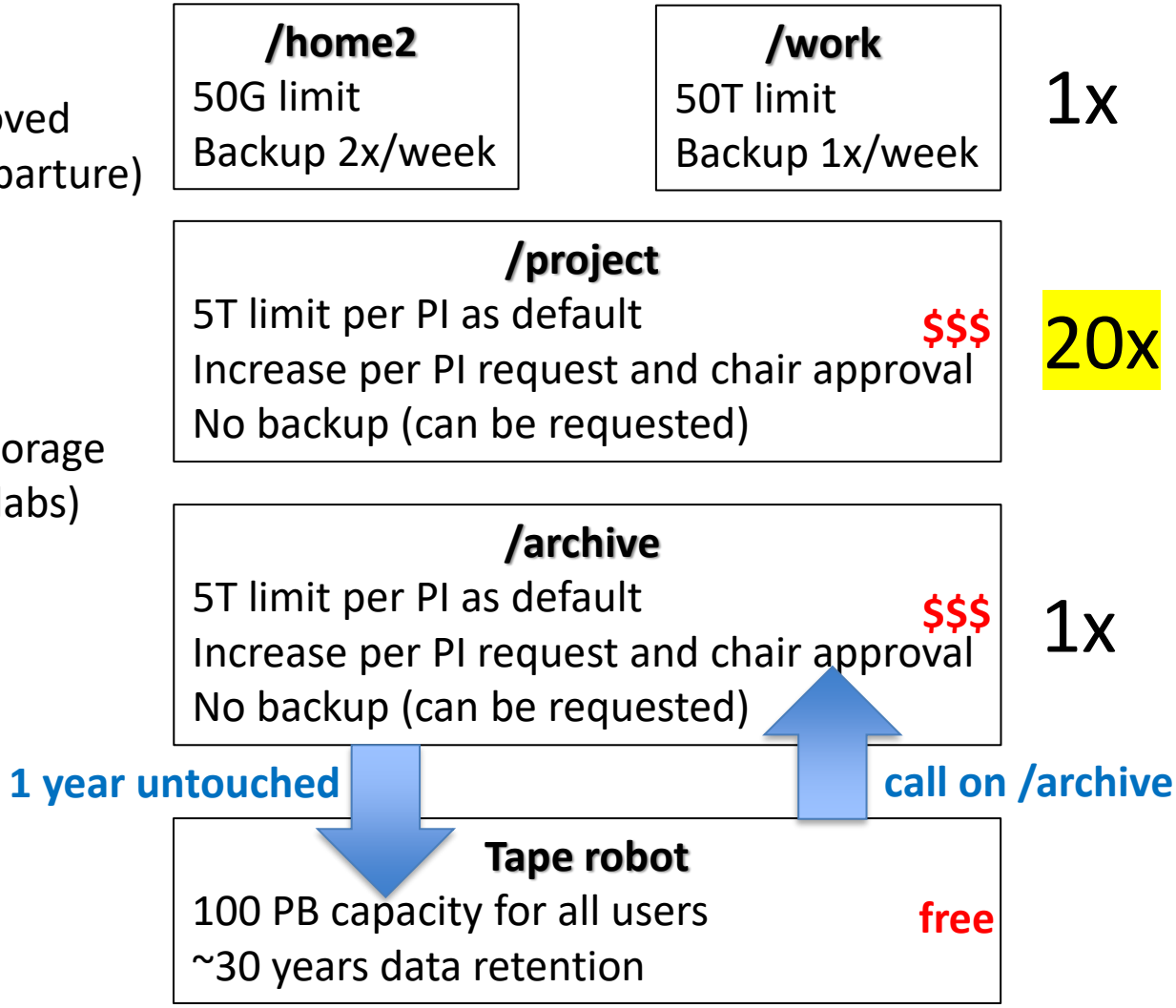
/archive
5T limit per PI as default **$$$**
Increase per PI request and chair approval
No backup (can be requested)

~ similar performance

Storage v2.0

User-centric
(will be removed with user departure)

Lab-centric
(long-term storage for research labs)



Storage v2.0 – new storage strategy

1. Build lab project structure on /archive
   - All data sharing approaches supported
   - Older data *automatically* transferred to tape robot
     - Identical name space for data on disk and tape
   - Significant storage cost reduction per PI
   - Decent I/O performance for single memory loads

2. Move data to /project for compute tasks with intensive, dynamic I/O
   - Temporary copy of raw data from /archive
   - Contained processing trees and results for easy deposition in project structure on /archive