Spring Forward into New Storage: FAS RC Data Storage Offerings

Sarah Marchese
Research Data Manager
FAS Research Computing

Florals? For spring? Groundbreaking.

Introduction

About Me

Sarah Marchese
Research Data Manager, FAS Research Computing

Research Data Management

- Collaborate with faculty, staff, and researchers to better understand, manage, classify, organize, and store research data throughout the data lifecycle
- Provide consultation and training on data storage, organization, and sharing
- Develop data management related resources and tools to track storage usage and prepare data for sharing and reuse
- Refine data transfer processes to migrate data to and from storage environments

FAS Research Computing

Research Computing Services:

- High-performance compute (HPC) cluster, Cannon
- Secure enclave for sensitive data (FASSE)
- Research storage (Active, Scratch, and Tape)
- Scientific software and applications
- Data science consultation
- Training seminars and workshops

Statistics:

- Manage 800+ lab groups and 7000+ accounts
- 76+ PiB of research storage across 3 data centers
- 99,900 CPU cores, 1000+ GPUs, and 1500+ compute nodes

Learning Objectives

- Data policies and principles
- Storage terminology and definitions
- Data storage workflow
- FASRC storage offerings
- Data storage tools
- Data retention and preservation
- Data security and privacy
- Additional storage options
  - Data repositories
- Data organization

Research Data Lifecycle

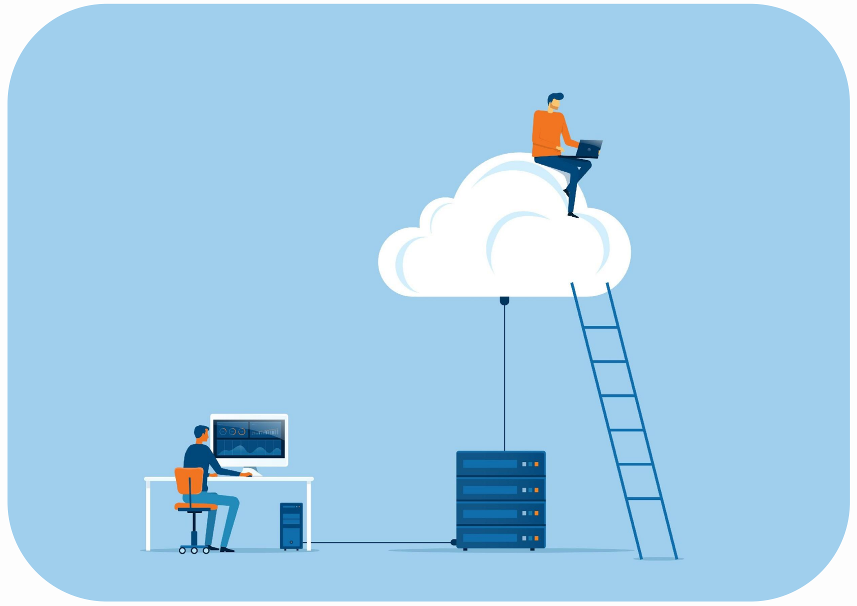
Planning

- Policies and procedures
- Data management plans (DMPs)
- Data Use Agreements (DUAs)
- Roles and responsibilities
- Data organization
- File naming conventions and directory structures

Creation & Analysis

- Collaborative tools
- Electronic Lab Notebooks
- Data transfer tools

Storage



- Active and long-term storage
- Data retention
- Storage Options
- Data security and privacy
- Data backups and prevention
- Data destruction and cleanup
- Storage tools

Sharing & Reuse

- Data repositories
- Open access data

Data Storage Policies

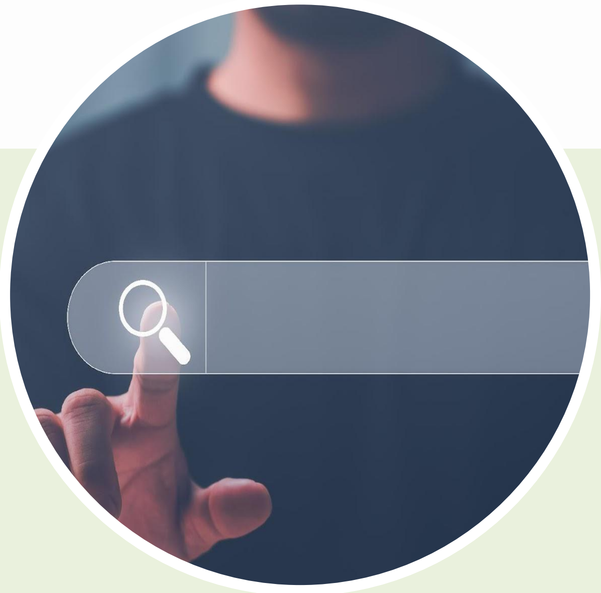
- **University policies:**
  - Research Data Ownership Policy
  - Harvard Research Data Security Policy (HRDSP)
  - Research Safety Application (Sensitive Research)
  - Retention and Maintenance of Research Records and Data Frequently Asked Questions ("FAQs"): "essential research records" need to be retained for a period of no fewer than seven (7) years after the end of a research project or activity.
  - Harvard University General Records Schedule

- **Funder requirements and policies:**
  - NIH Policy for Data Management and Sharing (2023)
  - NSF Data Management Plan Requirements and Data Sharing Policy

- **Additional policies:**
  - GDPR Research Guidance

FAIR Data Principles

The FAIR Data Principles published in Scientific Data in 2016 are a set of guiding principles proposed by scientists and organizations to encourage the reusability of digital research.

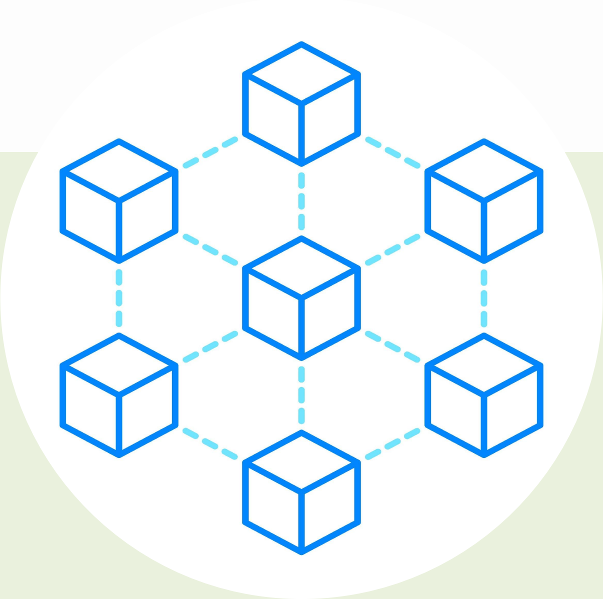


FINDABLE

Is your data discoverable by others?

ACCESSIBLE

Is your data available to others?



INTEROPERABLE

Can your data be integrated with other data?

REUSABLE

Can your data be reused by others?

Fair Data. Universiteit Gent. (2024, January 11). https://www.ugent.be/en/research/openscience/datamanagement/after-research/fair-data.htm

Data Storage Terminology

- **allocation**- Amount of storage space provided to an individual or group
- **archive**- The location and act of transferring materials to a facility authorized to appraise, preserve, and provide access to the data.
- **data retention**- Amount of time data is stored to adhere to policies or other requirements.
- **directory/path**- The folder or location on the computer where data is stored; the exact location of the file contents. Ex: /n/storage/lab_name/data/lab_member
- **Disaster Recovery**- Copy of an entire file system that can be used internally by FASRC in case of system-wide failure.
- **filesystem**- The server or machine where data is stored.
- **metadata**- Structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage the data
- **quotas**- The total amount or limit allowed for that storage folder.
- **repository**- A place to house, organize, and make data available for use
- **scratch**- temporary storage space for active data; often connected to a compute cluster
- **Snapshots**- Copies of a directory taken at a specific moment in time; a self-service recovery option for overwritten or deleted files within a specific time period.
- **tar/tarring**- A command line tool that bundles many files into a single file; helpful for moving data to Tape storage

Research Data at Harvard

- Resulting from **projects conducted at the University or on Harvard property**
  - Examples: In your lab, office, classroom, etc.
- Developed or collected under the auspices of the University, **even if research activities are occurring elsewhere**
  - Examples: Interviewing study participants in another country or utilizing data co-developed at a collaborator institution
- Developed or collected with **University resources (equipment, funding, etc.)**

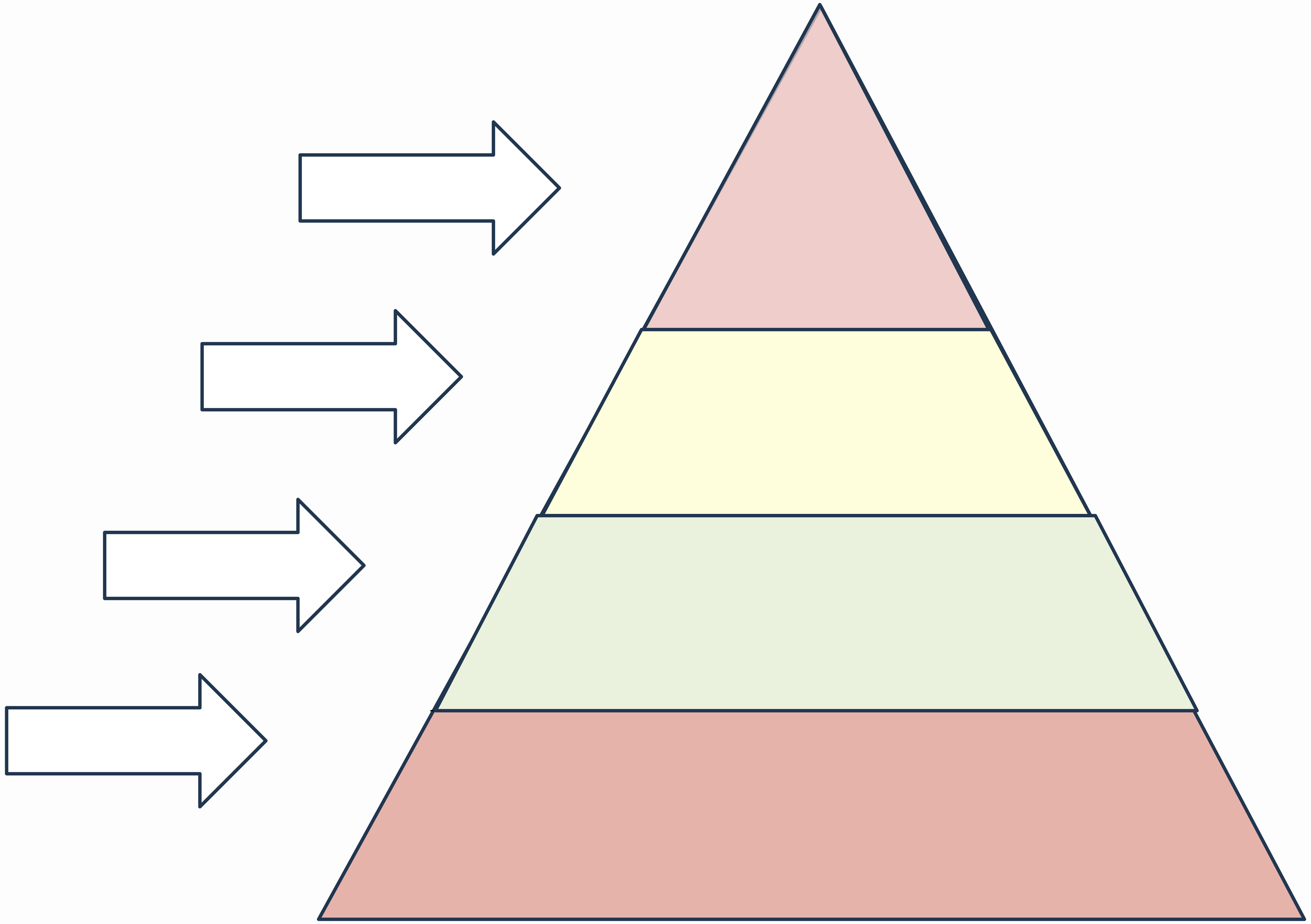
Types of Research Data

- **Published Data**
  How does the data support your research question?

- **Analyzed Data**
  What does the data tell us?

- **Processed Data**
  How can the raw data be manipulated?

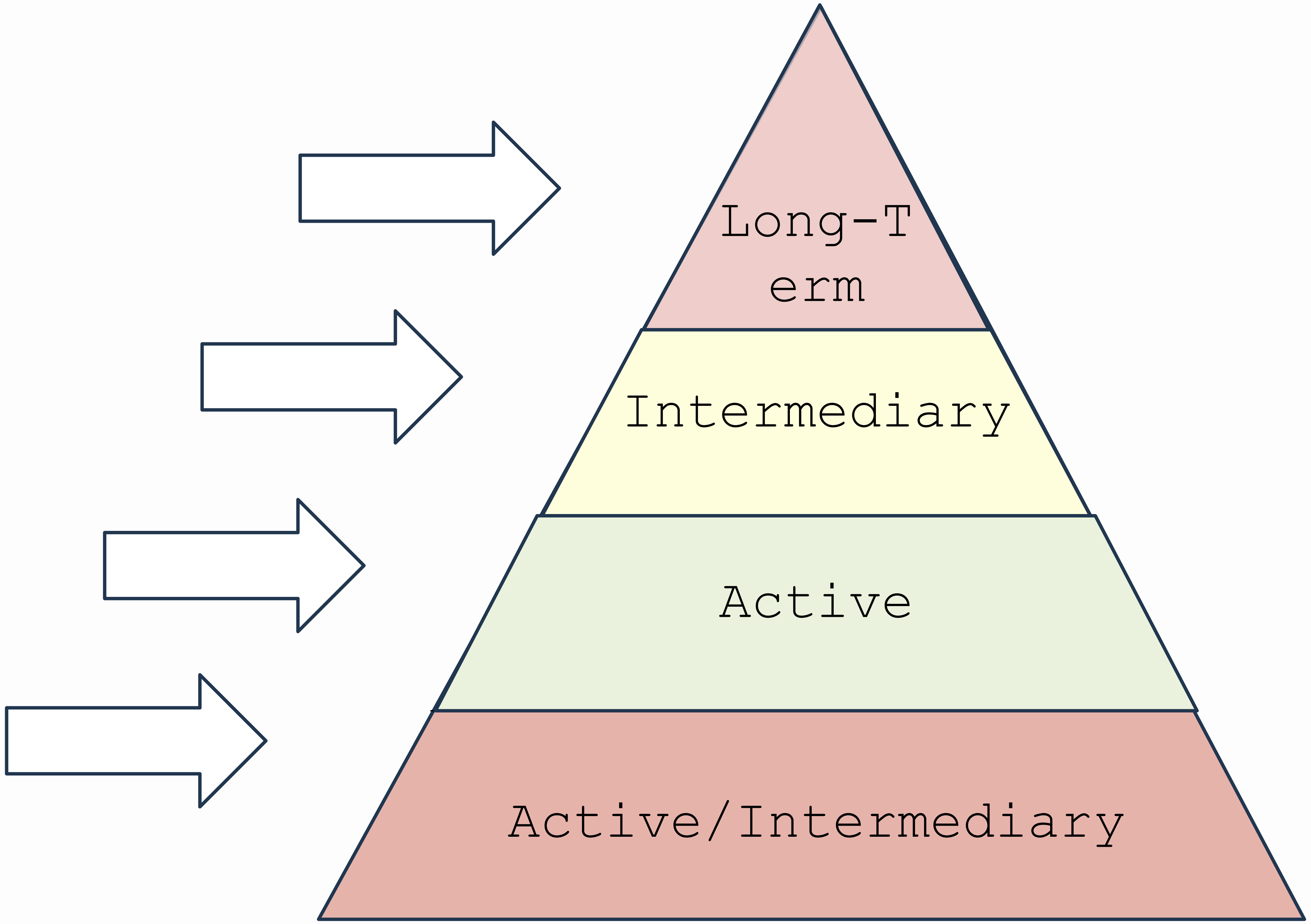
- **Raw Data**
  What is being measured or observed?



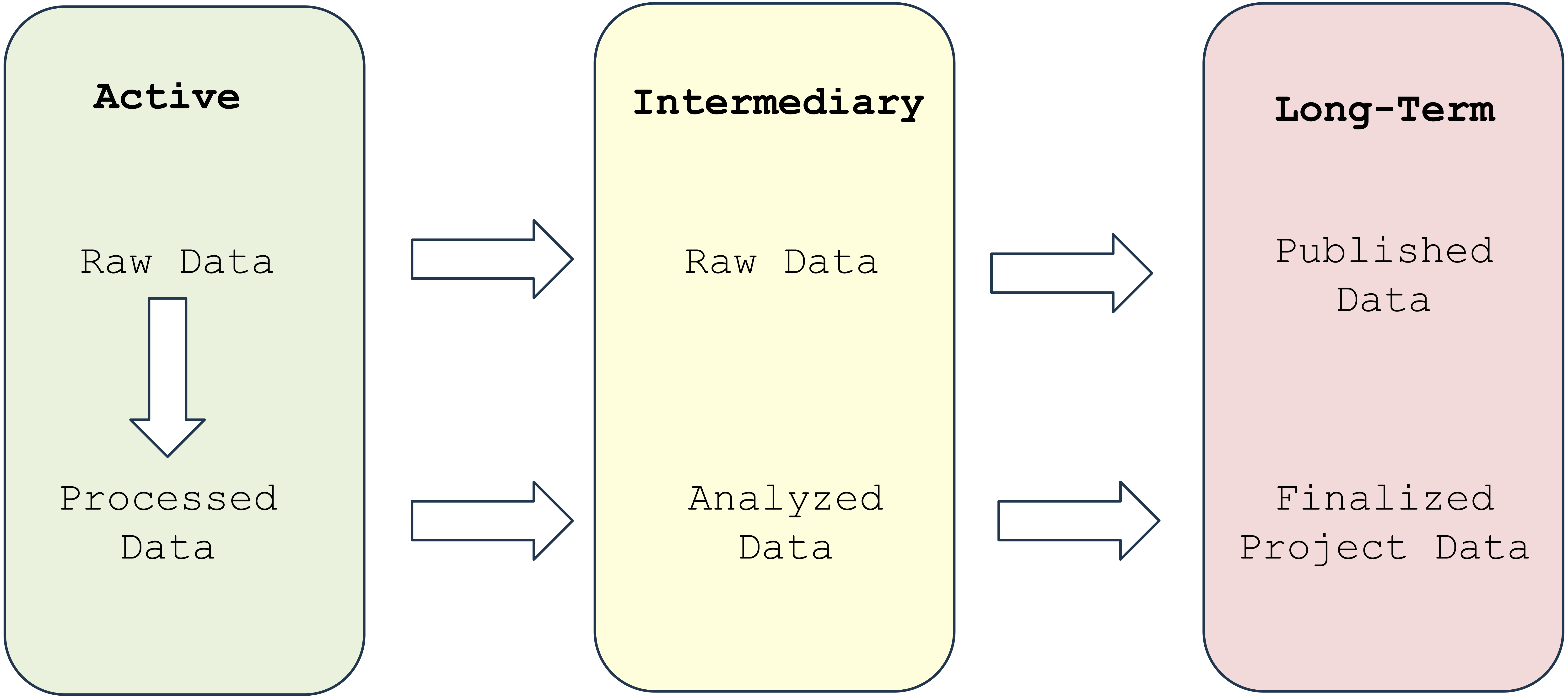
Types of Research Data

- **Published Data**
  How does the data support your research question?
- **Analyzed Data**
  What does the data tell us?
- **Processed Data**
  How can the raw data be manipulated?
- **Raw Data**
  What is being measured or observed?

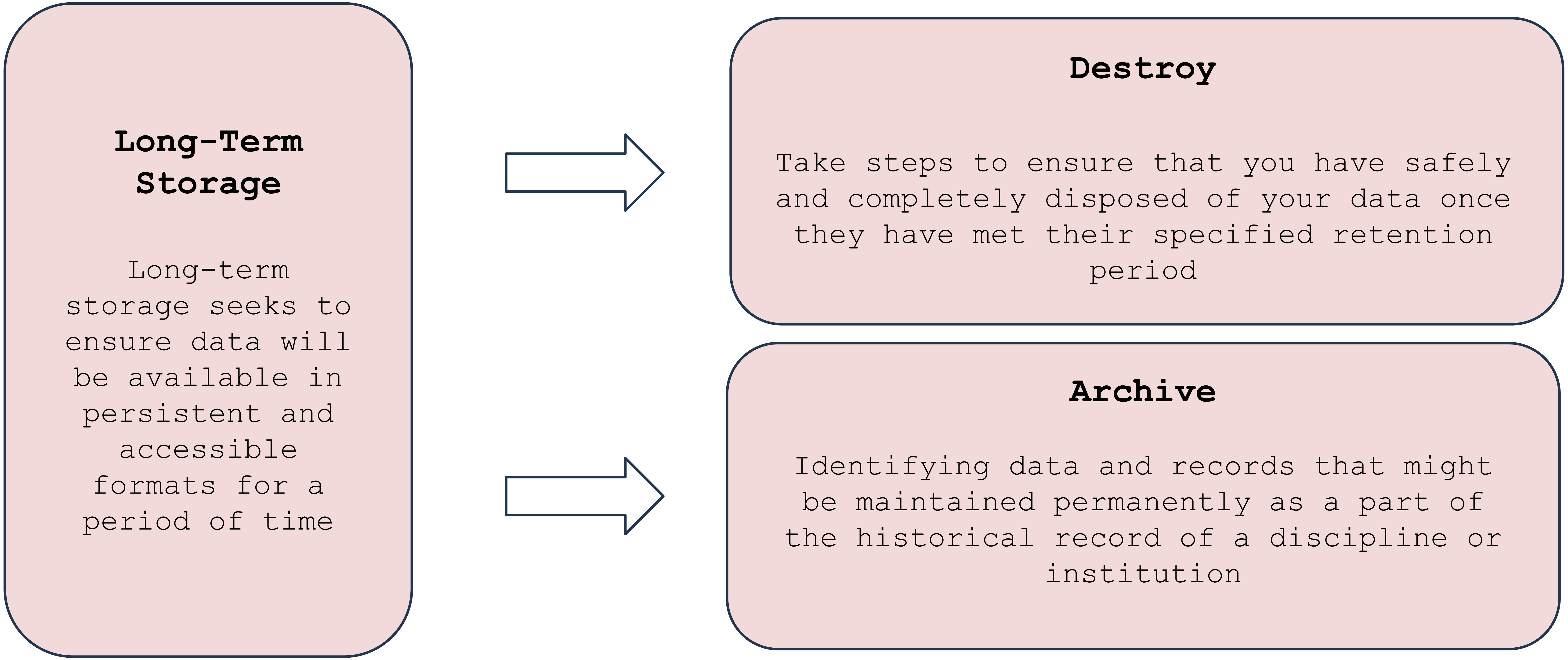
Data Storage



Data Storage Workflow



Data Storage Workflow



Long-Term Storage

Long-term storage seeks to ensure data will be available in persistent and accessible formats for a period of time

Destroy

Take steps to ensure that you have safely and completely disposed of your data once they have met their specified retention period

Archive

Identifying data and records that might be maintained permanently as a part of the historical record of a discipline or institution

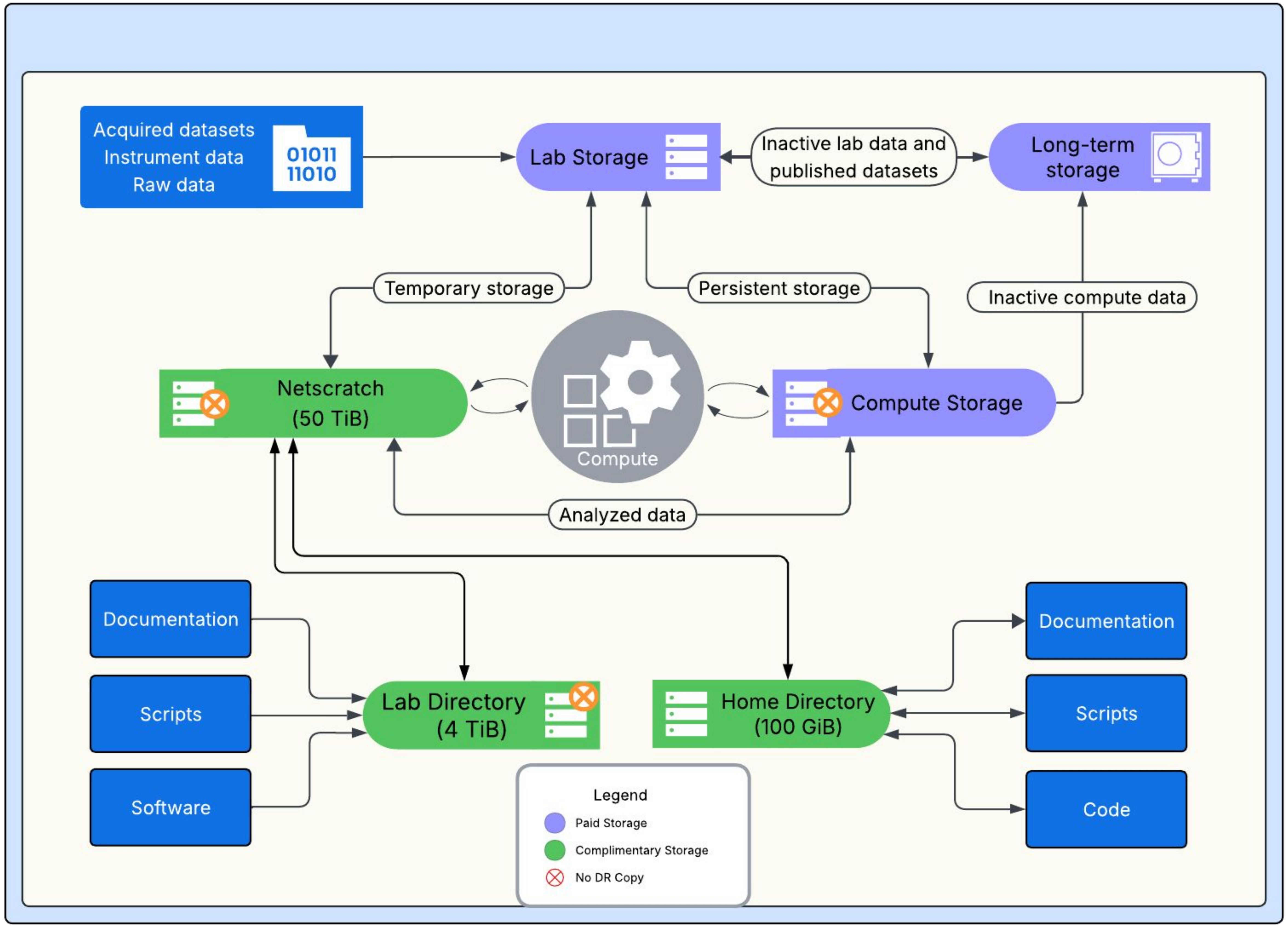
FASRC Storage Offerings (Complimentary)

	Home Directory	**Lab Directory**	**netscratch**
Description	Personal user storage. Not recommended for computational purposes.	General lab storage. Install software to be referenced from netscratch.	Temporary storage location for high performance data analysis.
Performance	Moderate	Moderate	High
Size	100GiB (fixed)	4TiB (fixed)	50TiB (fixed)
Mount	/n/homeNN/username	/n/holylabs	/n/netscratch
Retention	Daily snapshots weekly. Weekly snapshots every 4 weeks. Disaster recovery.	No snapshots. No disaster recovery.	No snapshots. No disaster recovery.
Cost	None	None	None
Security	Up to Level 2	Up to Level 2	Up to Level 2
Distribution	Folder generated for each user when granted cluster access. Limited to 100GiB.	Folder generated for each approved PI and their group. Limited to 4TiB.	Accessible to group members.

FASRC Storage Offerings (Paid)

	Compute Storage (Active)	Lab Storage (Active)	Long-term storage (Intermediary)	Tape (NESE) (Long-Term)	FASSE (Secure)
Description	Active storage for data analysis. Highly performant cluster adjacent storage. Optimized for AI/ML workflows.	General purpose storage for raw and project data. Can be used as buffer storage.	On-premise long-term storage option for Harvard affiliated labs.	Long-term storage of inactive research data. Externally managed.	Secure storage environment sensitive data; data generated using Data Use Agreements (DUAs) or IRB.
Performance	High	Moderate	Low	None	Moderate
Size	Available upon request	Available upon request	Available upon request	20TB increments.	Available upon request
Mount	/n/compute_storage/pi_lab	/n/lab_storage/pi_lab	/n/long_term/pi_lab	Transfer data to Tape using Globus	/n/fasse/pi_lab_projectname_l3
Retention	Weekly snapshots for 2 weeks. No disaster recovery.	Daily snapshots weekly. Weekly snapshots every 4 weeks. Disaster recovery.	No snapshots. Disaster recovery at additional cost.	No snapshots. No disaster recovery.	Daily snapshots weekly. Weekly snapshots every 4 weeks. Disaster recovery. Encryption at rest.
Cost	$150/yr per TiB	$125/yr per TiB	$30/yr per TiB	$15/yr per TB	$150/yr per TiB
Security Level	Level 2	Level 2	Level 2 (Up to Level 3)	Level 2	Up to Level 3

Data Storage Workflow



Acquired datasets
Instrument data
Raw data

01011
11010

Lab Storage

Inactive lab data and published datasets

Long-term storage

Temporary storage

Persistent storage

Inactive compute data

Netscratch (50 TiB)

Compute

Compute Storage

Analyzed data

Documentation

Scripts

Software

Lab Directory (4 TiB)

Home Directory (100 GiB)

Documentation

Scripts

Code

Legend

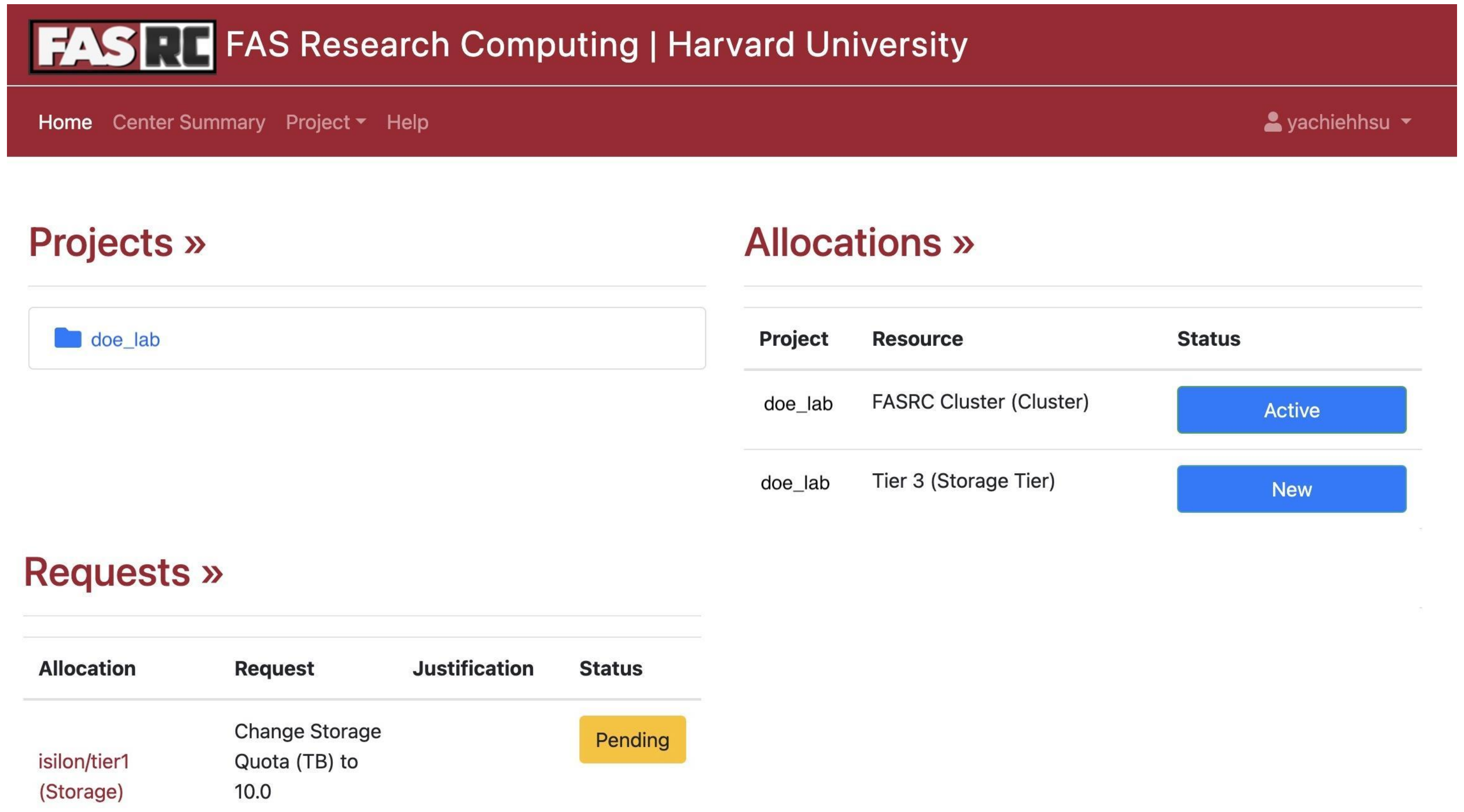
- ● Paid Storage
- ● Complimentary Storage
- ⊗ No DR Copy

FAS RC Storage

- Every FAS RC group is provided with two storage locations as a default
  - Lab Directory (/n/holylabs)
  - Netscratch (/n/netscratch)
- Storage folders contain two subdirectories as a default
  - Lab- Accessible by all lab members and viewable in Globus
  - Everyone- Accessible by anyone on the compute cluster, ideal for cross-group collaboration
- Storage can be accessed via the command line interface (CLI), Open OnDemand (OOD) or Globus
- FASRC quota command provides storage limit and usage information for all FASRC storage options except Tape
  - quota <PATH>

Storage Tools: Coldfront

- Open-source resource allocation management system
- Enables viewing and management of lab groups, storage and cluster allocations
  - View/add projects (lab groups)
  - View/add/remove users
  - Adjust notifications
  - Request new storage allocations
  - Request changes to existing storage allocations
  - Edit user roles (assign manager status)



Projects »

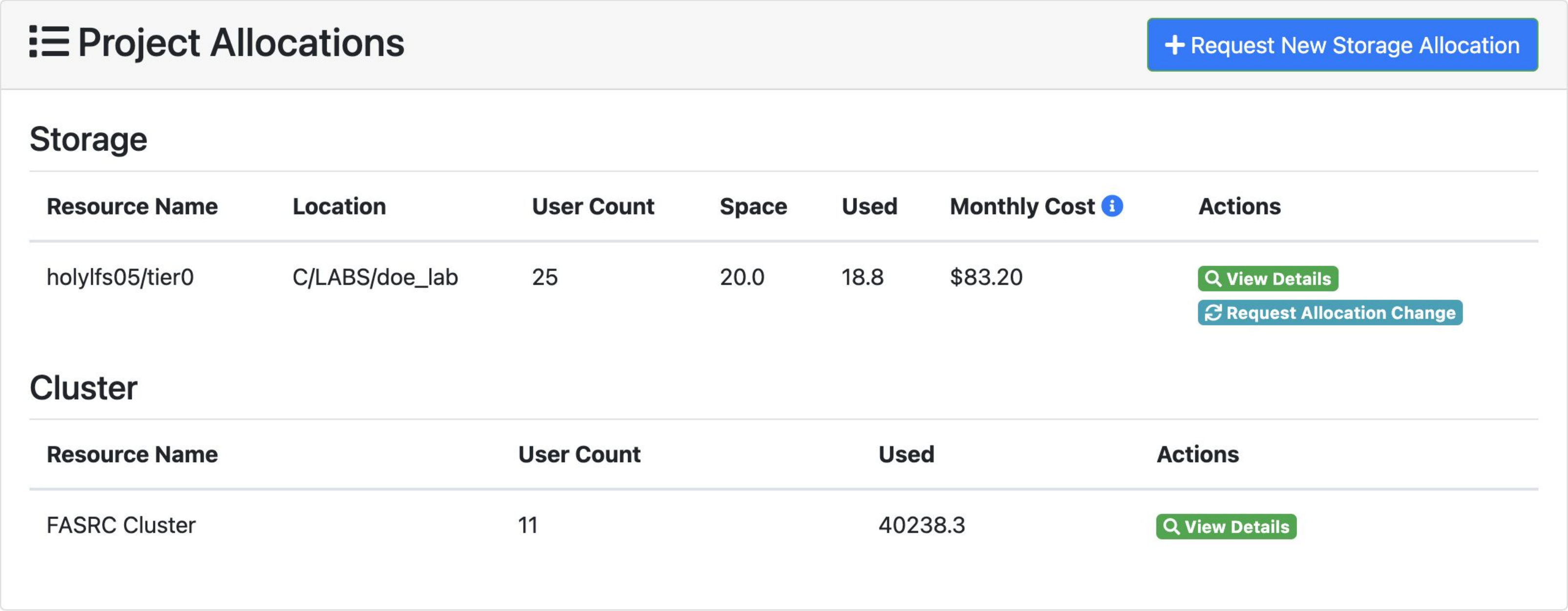
doe_lab

Allocations »

Project	Resource	Status
doe_lab	FASRC Cluster (Cluster)	Active
doe_lab	Tier 3 (Storage Tier)	New

Requests »

Allocation	Request	Justification	Status
isilon/tier1 (Storage)	Change Storage Quota (TB) to 10.0		Pending



Project Allocations

Storage

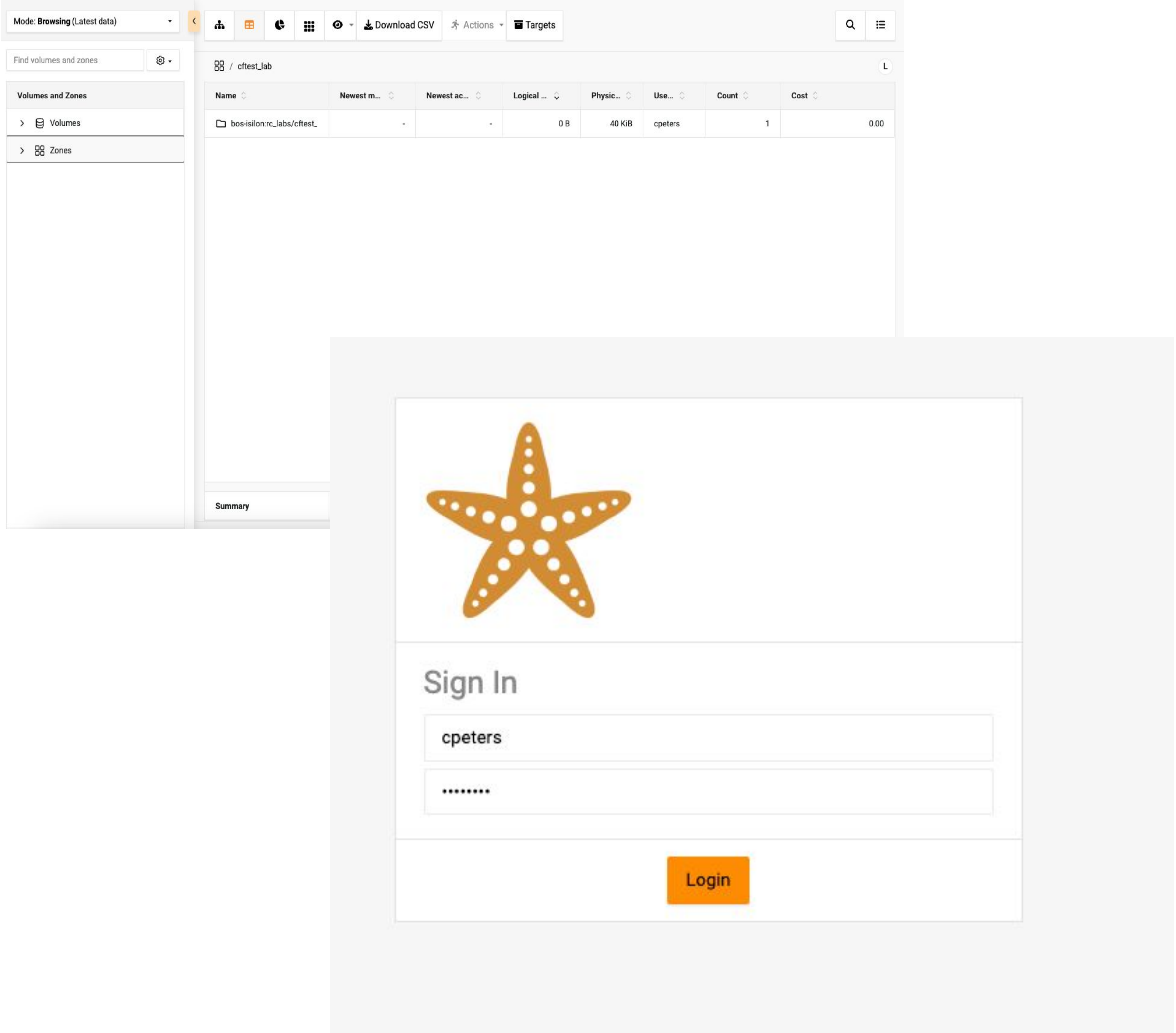
Resource Name	Location	User Count	Space	Used	Monthly Cost	Actions
holylfs05/tier0	C/LABS/doe_lab	25	20.0	18.8	$83.20	View Details Request Allocation Change

Cluster

Resource Name	User Count	Used	Actions
FASRC Cluster	11	40238.3	View Details

Storage Tools: Starfish Zones

- Self-service visual tool enabling users to view group storage amounts and locations
- Navigate folder structures to access detailed information about files and storage
- Utilize the tool to assist with data organization and cleanup efforts, including key information about the group or lab’s usage over time
- Information can be exported to CSV



Mode: **Browsing** (Latest data)

cftest_lab

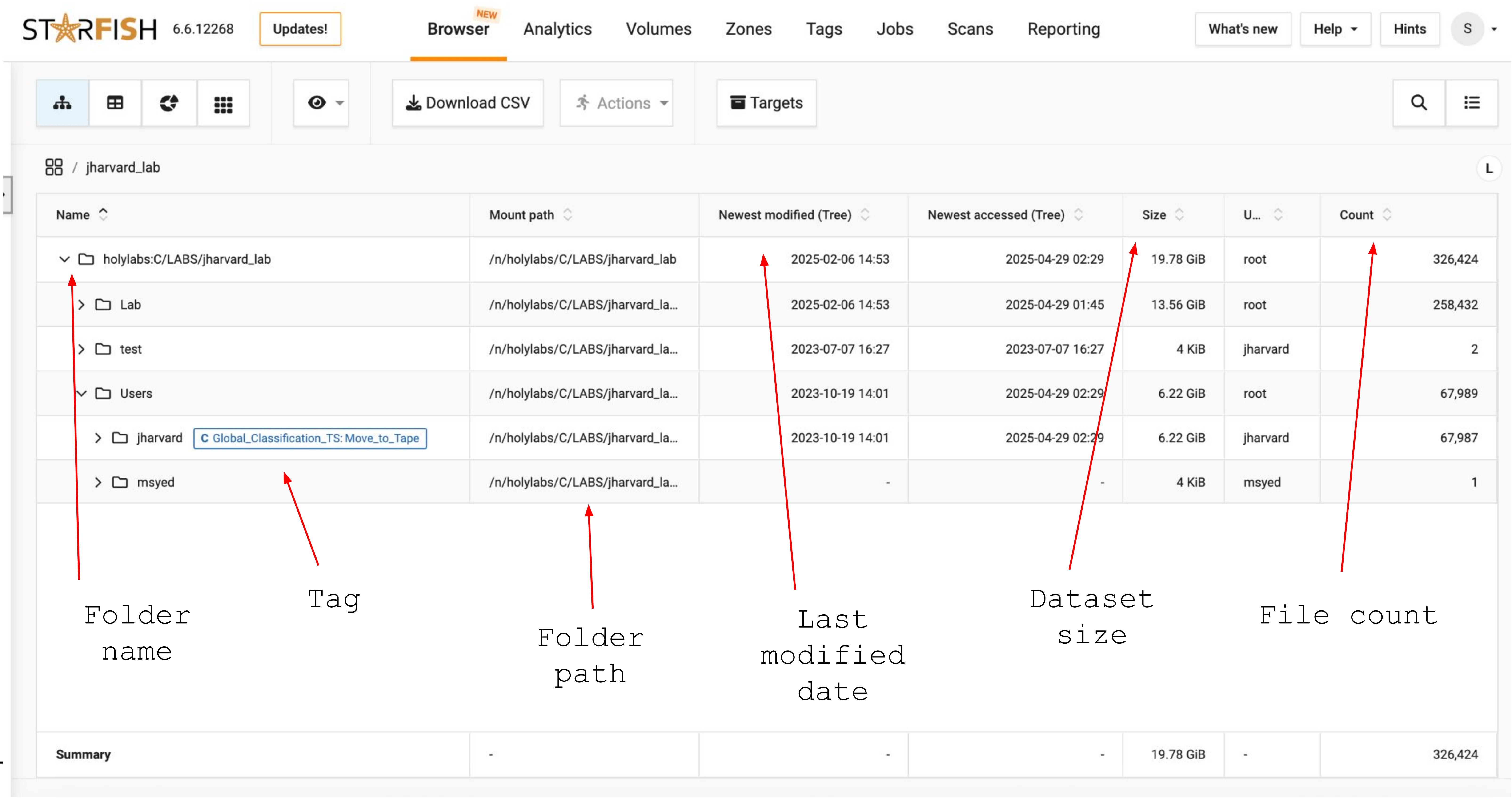
Name	Newest m...	Newest ac...	Logical ...	Physic...	Use...	Count	Cost
bos-isilon:rc_labs/cftest_	-	-	0 B	40 KiB	cpeters	1	0.00

Sign In

cpeters

••••••••

Storage Tools: Starfish Zones



jharvard_lab

Name	Mount path	Newest modified (Tree)	Newest accessed (Tree)	Size	U...	Count
holylabs:C/LABS/jharvard_lab	/n/holylabs/C/LABS/jharvard_lab	2025-02-06 14:53	2025-04-29 02:29	19.78 GiB	root	326,424
Lab	/n/holylabs/C/LABS/jharvard_la...	2025-02-06 14:53	2025-04-29 01:45	13.56 GiB	root	258,432
test	/n/holylabs/C/LABS/jharvard_la...	2023-07-07 16:27	2023-07-07 16:27	4 KiB	jharvard	2
Users	/n/holylabs/C/LABS/jharvard_la...	2023-10-19 14:01	2025-04-29 02:29	6.22 GiB	root	67,989
jharvard C Global_Classification_TS: Move_to_Tape	/n/holylabs/C/LABS/jharvard_la...	2023-10-19 14:01	2025-04-29 02:29	6.22 GiB	jharvard	67,987
msyed	/n/holylabs/C/LABS/jharvard_la...	-	-	4 KiB	msyed	1
Summary	-	-	-	19.78 GiB	-	326,424

Folder name

Tag

Folder path

Last modified date

Dataset size

File count

Data Retention

Research records should generally be retained no fewer than seven (7) years after the end of a research project or activity (Harvard data retention policy)

Evaluate for Retention

- Identify & retain “essential research records”.
- “Essential” Research Records are:
    - Records associated with grant applications, proposals, and other funding requests
    - Records needed to substantiate compliance with sponsored research
    - Records associated with published research and patents
    - Scholarship considered for long-term preservation and access by the University Archives or the local archives of the Schools
    - Data or materials designated as essential by the Schools and relevant disciplines
- Organize and annotate appropriately

Retention Policies:

- Retention and Maintenance of Research Records and Data Frequently Asked Questions (FAQ)
- Harvard University General Records Schedule (GRS)

Tape Storage

- Data can be copied to Tape, and retrieved from Tape using the Globus tool
- Storage allocations are provided in 20TB tapes
- Size limitation:
  - 10,000 files per directory
  - File sizes 1-100GB
- Data that does not meet the Tape restrictions needs to be tarred prior to migration
- No direct access available, metadata provided by Globus
- Cost: $15/yr per TB
- Security level: Up to Harvard Data Security Level 2

FASSE

- FAS Secure Enclave (FASSE) is a secure storage environment for data analysis or sensitive data
  - Data generated using Data Use Agreements (DUAs) or IRB
- Harvard Data Security Level 3
- Data in FASSE should not be transferred to/from Cannon; they are different security levels
- Does not allow direct access to the internet; requires connection to FASRC VPN
- Request process differs from other FASRC storage; PIs need to follow the HRDSP and Associated Guidance
  - FASRC will review relevant documents (DUA/DAT/IRB)
  - Fill out FASSE New Project Request Form

Data Use Agreements

What is a Data Use Agreement?

- The transfer of confidential, proprietary or sensitive data between organizations requires a **formalized written agreement or contract between the two organizations.**
- The written contract, or Data Use Agreement (DUA) will outline the **terms and conditions of the data transfer.**

How to Comply:

- DUAs must be reviewed and signed by the Office for Sponsored Programs
- The project PI or group leader is responsible for ensuring access to the data is compliant with the DUA
- The DUA Guidance and Policy provides step-by-step instructions for researchers on the procedures for submitting and managing DUA requests in the Agreement System

Why are DUAs important?

- They help to avoid misunderstandings and disputes over the use and storage of data, access and security measures, and other important factors, including publication rights and ownership of results

Data Security and Privacy

- Required to **protect the privacy of research subjects and to secure sensitive and personally identifiable information (PII)**
- Properly protecting research data is a fundamental obligation grounded in the values of stewardship, integrity, and commitments to the providers and sources of the data
- The **University's Intellectual Property (IP)** policy governs the ownership and disposition of IP including, but not limited to, inventions, copyrights (including computer software), trademarks, and tangible research property such as biological materials
- Harvard maintains a multi-level **security system from Level 1-5**

Harvard Data Security Levels

Level 1 - Publicly available and unrestricted data
Storage: Public repositories, consumer products

Level 2 - Unpublished non-sensitive research data
Storage: Harvard standard email

Level 3 - **Sensitive Data** and some regulated data that could be damaging
Storage: Harvard Dropbox, Shared network, OneDrive, SharePoint

Level 4 - **Sensitive Data** that could place the subject at significant risk
Storage: Harvard Secure Transfer, External hard disk with encryption

Level 5 - **Sensitive Data** that could place the subject at severe risk of harm
Storage: Requires security consulting for special handling

Harvard Storage Tools: Security Levels

	Level 1	**Level 2**	**Level 3**	**Level 4**
Consumer Google Drive - All tools	✔	*	❌	❌
Consumer Dropbox, Evernote	✔	*	❌	❌
Harvard email (M365, Gmail)	✔	✔	❌	❌
Harvard Confluence/Wiki	✔	✔[1]	✔[1]	❌
Harvard GitHub code.harvard.edu	✔	✔[1]	✔[1]	❌
Harvard Dropbox	✔	✔[1]	✔[1]	❌
Harvard Google Drive/Docs (g.harvard)	✔	✔[1]	✔[1]	❌
Harvard Slack	✔	✔	✔	❌
Harvard M365 (OneDrive, SharePoint, Teams)	✔	✔	✔	❌
Harvard M365 SharePoint with L4 configuration	✔	✔	✔	✔[2]
Harvard Qualtrics with L4 configuration	✔	✔	✔	✔[2]

Data Security: Backups and Prevention

2-2-1 Rule: Two copies, two storage formats, with one type offsite

2 copies

2 storage formats

1 off-site

Crashplan Software: Ensures critical data is recoverable in the event of data loss or deletion

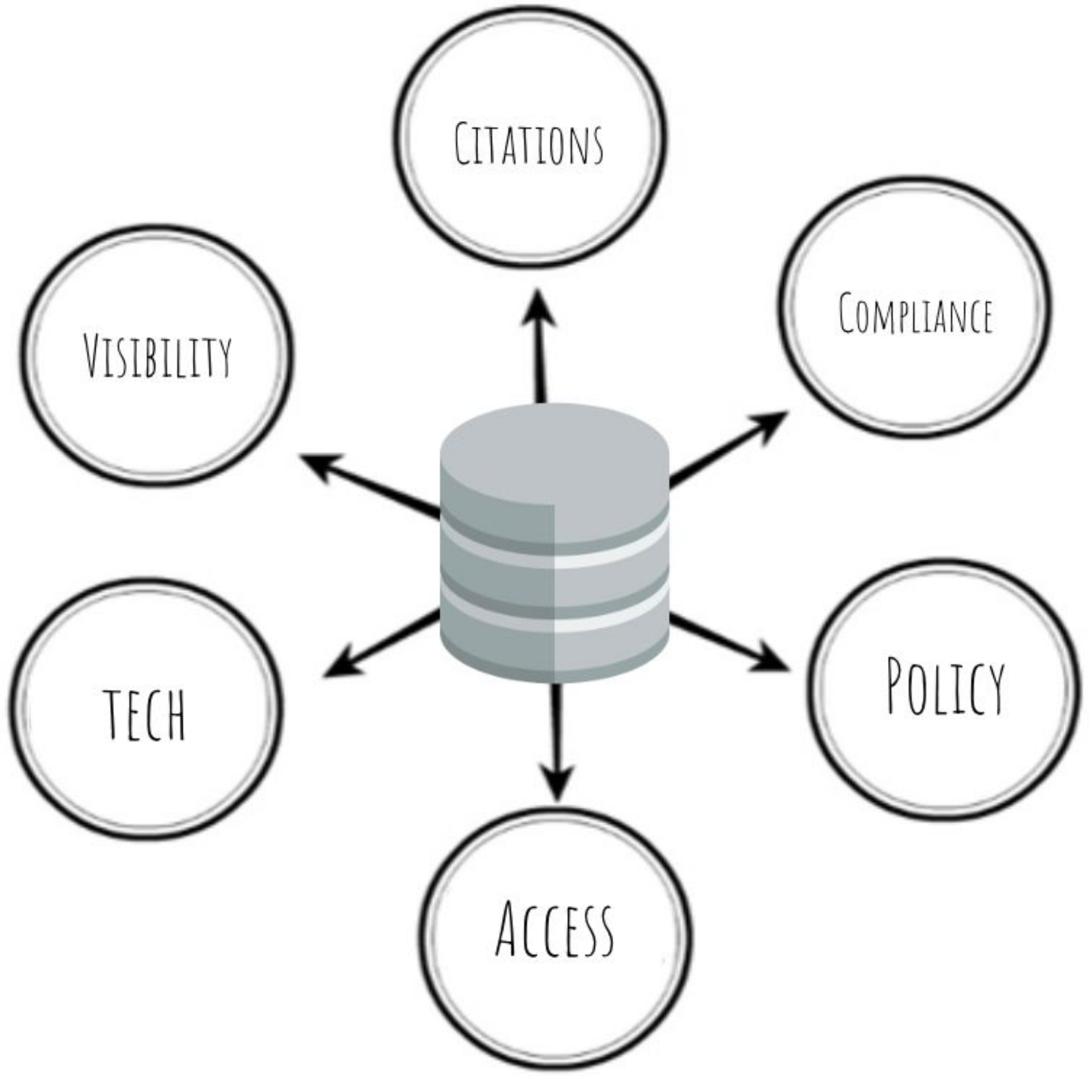
- Backs up continually over almost any network on or off-campus
- Recovers documents from any computer via a web browser
- Stores document copies for a minimum of 60 days

Additional Storage Options

	Electronic Lab Notebook (ELN): **RSpace**	**Project Management:** **Open Science Framework (OSF)**	**Code repository:** **GitHub**
Description	• Open-source tool supported by University Research Computing (URC) • Helps researchers organize, store, and share protocols, analysis, and experimental notes in a centralized and secure platform	• A free and open-source project management tool that supports researchers throughout the project lifecycle	• Web-based service for Git repositories • Commonly used for managing and sharing versions of code for programming projects
Eligibility	• Available for free to faculty with a Harvard appointment • Login with HarvardKey authentication	• Available to users with a Harvard email address • Login with HarvardKey authentication	• Open-source tool, not hosted by Harvard
Features	• Collaborate across groups • Simplify data inventory and sample management • Integrate with popular research tools • Link to university supported data storage • Delegate administration of group access • Open and restricted data sharing • Export data in various formats	• Open and restricted data sharing • Upload datasets, documents, presentations, etc. and receive a unique identifier (DOI) for each item • Connects to popular research tools • Recognized by major funding bodies as a data repository for sharing research materials	• Effective version control tool for files and text documents • Large open-source community of users • Collaborative environment for updating code • Retain a copy of the files after project close, so they are available to the university

Data Repositories

- Repositories provide the technical infrastructure to store data, share data publicly and organize data in a logical way
- Supply a persistent identifier and a citation for your data
- Provide access controls (open or restricted)
- Compliant with funders and journals requirements
- Facilitate discovery of your data with search capabilities
- Preserve data on a long-term basis



Data Repositories

Institutional

HARVARD
Dataverse

NOAA
Institutional
Repository

UMass Chan
MEDICAL SCHOOL
eScholarship@UMassChan

Disciplinary

db GaP
GENOTYPES and PHENOTYPES

GEO
Gene Expression Omnibus

NIMH
National Institute
of Mental Health

Generalist

The
Dataverse®
Project

zenodo

figshare

Generalist Repositories

Beneficial characteristics of generalist repositories:

- Unique and persistent identifiers
- Long-term sustainability of datasets
- Metadata schemas
- Dataset curation and quality assurance
- Free and easy access to open data
- Data security and access controls
- Common formats
- Data retention policies
- Support FAIR data

The Dataverse® Project DRYAD figshare OSF

MENDELEY DATA Vivli zenodo

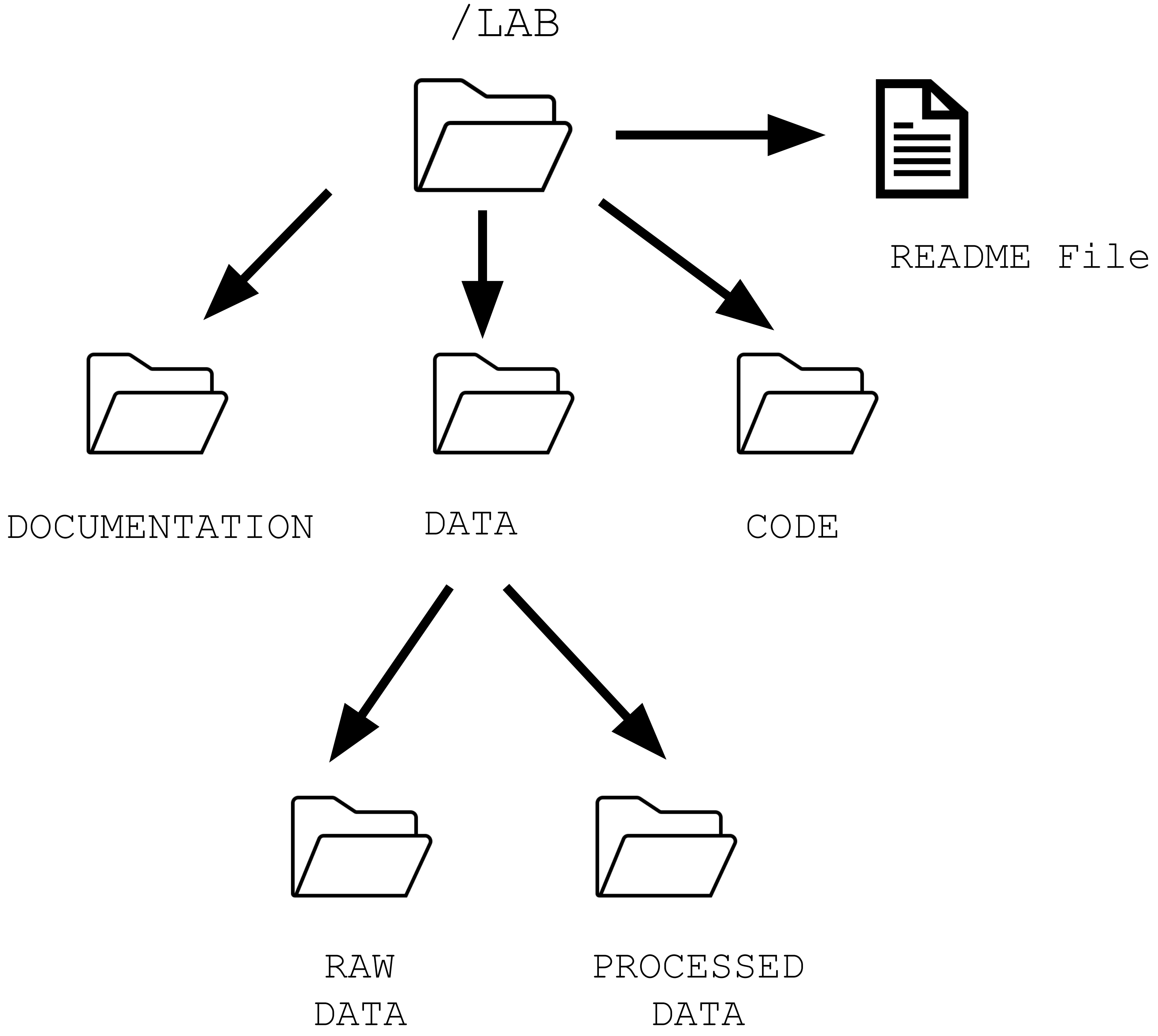
Harvard Dataverse and DASH

- **Harvard Dataverse:** Generalist data repository; open-source
  - Open to researchers from any discipline
  - Extended support for Harvard researchers
  - Share, archive, cite, and access research data
  - Paid data curation services offered
    - **Harvard users receive 2.5TB per account for free; maximum file size 2.5GB**
  - Option for large data storage (fee based Tape)
  - Sensitive data not supported
    - Data must be de-identified prior to deposit

- **DASH:** Harvard's central, open-access repository for archiving and sharing manuscripts
  - Managed by Harvard Library’s Office for Scholarly Communication (OSC)
  - Articles are free to download; available to everyone, free from most copyright and licensing restrictions
  - Supports browsing and search capabilities
    - Contents discoverable by search engines and HOLLIS

Data Organization: Directory Structure

- Arrange folders and files hierarchically
- One project, one folder
- Limit the number of files to a few thousand per folder
- Create “shallow” directories
  - Not too many nested folders
- Store and organize data based on the desired usage
- Represent the structure of information
  - Keep raw data and processed data separate
- Include a README file in the project folder for reference

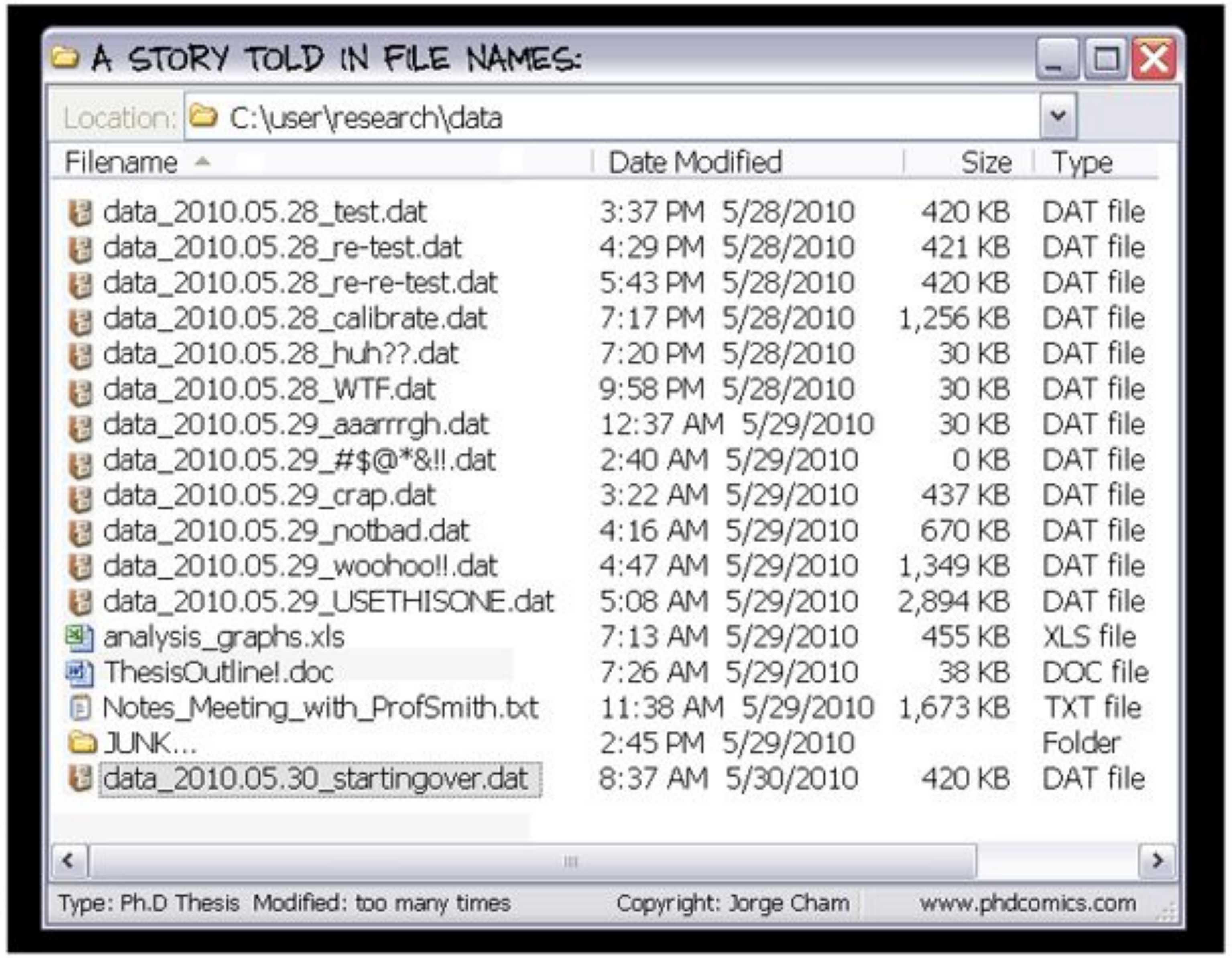


Data Organization: File Naming

- Establish consistent file naming conventions across the group or lab
- Describe what the files contain and how they relate to one another
- Include essential information, such as date, project title, and a unique identifier
- Use versioning to indicate the most current version of a document
- Avoid special characters and spaces (limit to 25 characters per name)
- Machine-readable file names preferred

Good Examples:

- Date_ExperimentName_InstrumentName_CaptureTime_ImageID.tif
- Date_ProjectName_DocumentName_v2.txt



A STORY TOLD IN FILE NAMES:

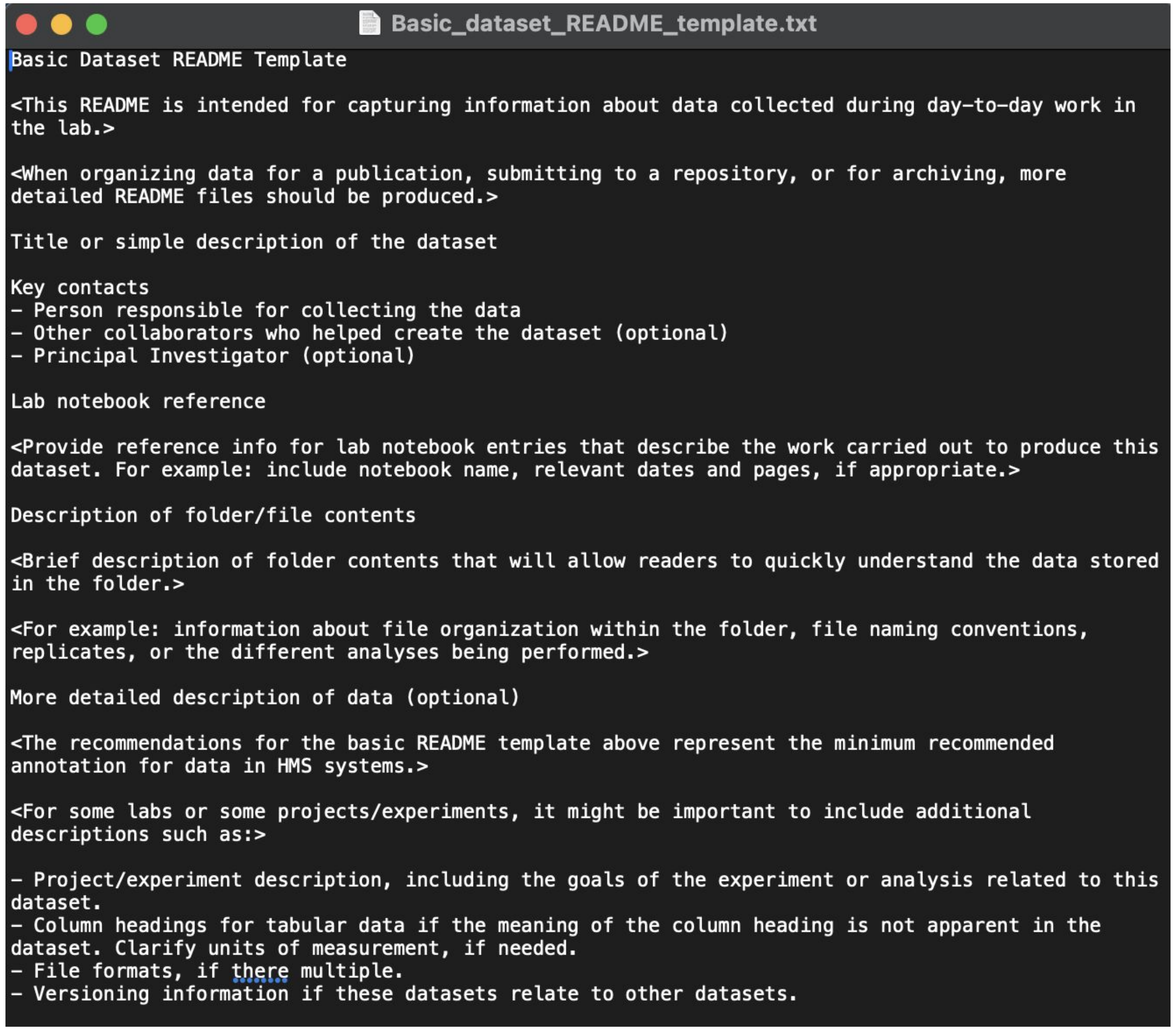
Location: C:\user\research\data

Filename	Date Modified	Size	Type
data_2010.05.28_test.dat	3:37 PM 5/28/2010	420 KB	DAT file
data_2010.05.28_re-test.dat	4:29 PM 5/28/2010	421 KB	DAT file
data_2010.05.28_re-re-test.dat	5:43 PM 5/28/2010	420 KB	DAT file
data_2010.05.28_calibrate.dat	7:17 PM 5/28/2010	1,256 KB	DAT file
data_2010.05.28_huh??.dat	7:20 PM 5/28/2010	30 KB	DAT file
data_2010.05.28_WTF.dat	9:58 PM 5/28/2010	30 KB	DAT file
data_2010.05.29_aaarrgh.dat	12:37 AM 5/29/2010	30 KB	DAT file
data_2010.05.29_#$@*&!!.dat	2:40 AM 5/29/2010	0 KB	DAT file
data_2010.05.29_crap.dat	3:22 AM 5/29/2010	437 KB	DAT file
data_2010.05.29_notbad.dat	4:16 AM 5/29/2010	670 KB	DAT file
data_2010.05.29_woohoo!!.dat	4:47 AM 5/29/2010	1,349 KB	DAT file
data_2010.05.29_USETHISONE.dat	5:08 AM 5/29/2010	2,894 KB	DAT file
analysis_graphs.xls	7:13 AM 5/29/2010	455 KB	XLS file
ThesisOutline!.doc	7:26 AM 5/29/2010	38 KB	DOC file
Notes_Meeting_with_ProfSmith.txt	11:38 AM 5/29/2010	1,673 KB	TXT file
JUNK...	2:45 PM 5/29/2010		Folder
data_2010.05.30_startingover.dat	8:37 AM 5/30/2010	420 KB	DAT file

Type: Ph.D Thesis Modified: too many times Copyright: Jorge Cham www.phdcomics.com

Data Organization: README File

- Record information necessary to understand the content and context of the data (directory structure, file naming convention, abbreviations etc.)
- Store this information in a README file alongside your research data
- Documentation is an ongoing process and should occur throughout the length of a project
- Write the README file as a plain text document



```
Basic Dataset README Template

<This README is intended for capturing information about data collected during day-to-day work in
the lab.>

<When organizing data for a publication, submitting to a repository, or for archiving, more
detailed README files should be produced.>

Title or simple description of the dataset

Key contacts
- Person responsible for collecting the data
- Other collaborators who helped create the dataset (optional)
- Principal Investigator (optional)

Lab notebook reference

<Provide reference info for lab notebook entries that describe the work carried out to produce this
dataset. For example: include notebook name, relevant dates and pages, if appropriate.>

Description of folder/file contents

<Brief description of folder contents that will allow readers to quickly understand the data stored
in the folder.>

<For example: information about file organization within the folder, file naming conventions,
replicates, or the different analyses being performed.>

More detailed description of data (optional)

<The recommendations for the basic README template above represent the minimum recommended
annotation for data in HMS systems.>

<For some labs or some projects/experiments, it might be important to include additional
descriptions such as:>

- Project/experiment description, including the goals of the experiment or analysis related to this
dataset.
- Column headings for tabular data if the meaning of the column heading is not apparent in the
dataset. Clarify units of measurement, if needed.
- File formats, if there multiple.
- Versioning information if these datasets relate to other datasets.
```

Harvard Longwood Medical Area Research Data Management Working Group. “Basic_Dataset_Readme_Template.Txt". 2026-02-04. https://osf.io/pw7ed/files/f862v

Storage Summary

- Review and adhere to data storage policies and procedures (institutional and funder)
- Develop and streamline a data storage workflow, including FASRC’s new data storage offerings
- Data storage tools can assist with data review and cleanup efforts, requesting new storage allocations, and modifying group membership
- Select an appropriate storage option based on the data retention and security requirements
- Alternative storage options are available across Harvard with various security levels
- Investigate data repositories for sharing
- Adopt data organization techniques to make your data discoverable and provide context

Please complete the seminar survey!

https://tinyurl.com/FASRC-training

Contact

rdm@rc.fas.harvard.edu

rchelp@rc.fas.harvard.edu

www.rc.fas.harvard.edu

www.rc.fas.harvard.edu/services/research-data-management/