Choosing data storage

@ Northwestern University

Northwestern
INFORMATION TECHNOLOGY

Instructor: Llewellyn Fernandes – Data Management Specialist

Date: January 15, 2025

Materials: https://github.com/nuitrcs/rdm-workshops

Poll
Where do you store your research data?

Today we'll cover

- Types of data storage
- Northwestern-provided options
- How to choose
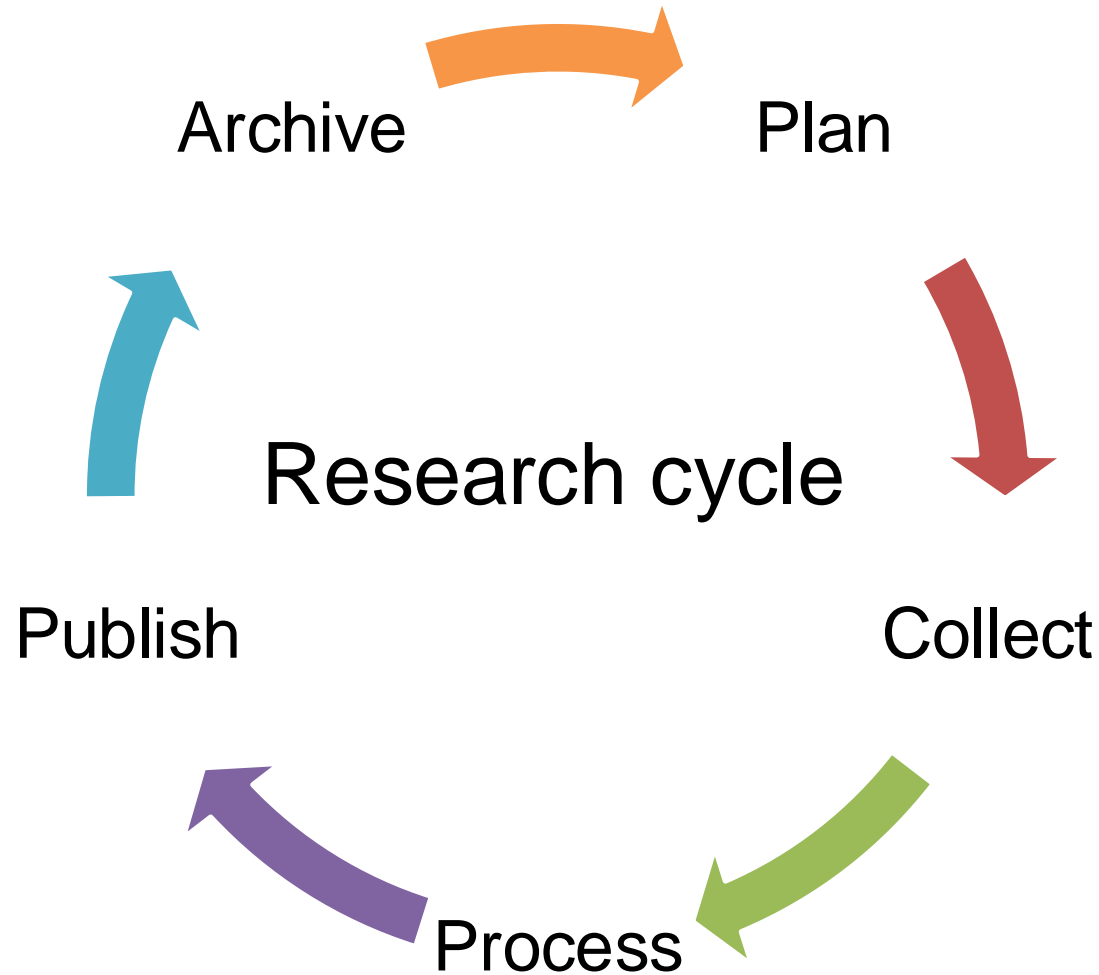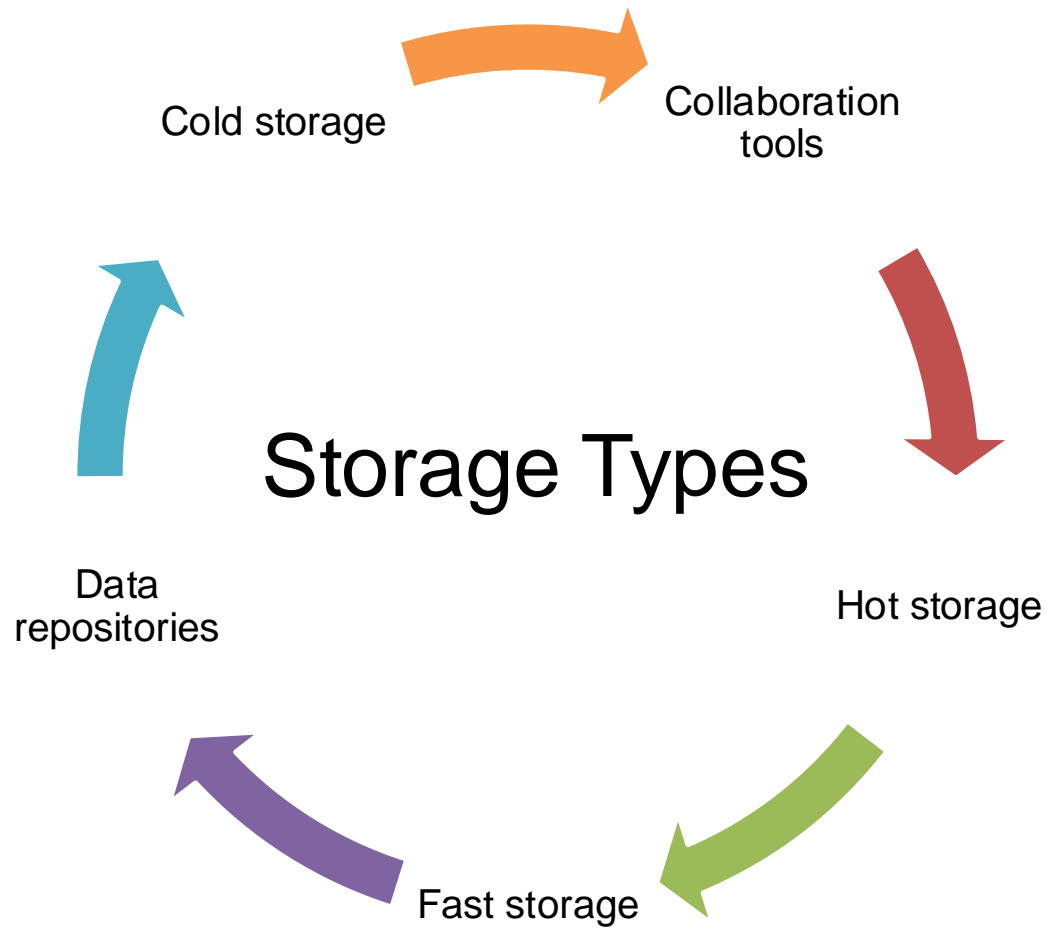- Planning your data workflow

tl;dr

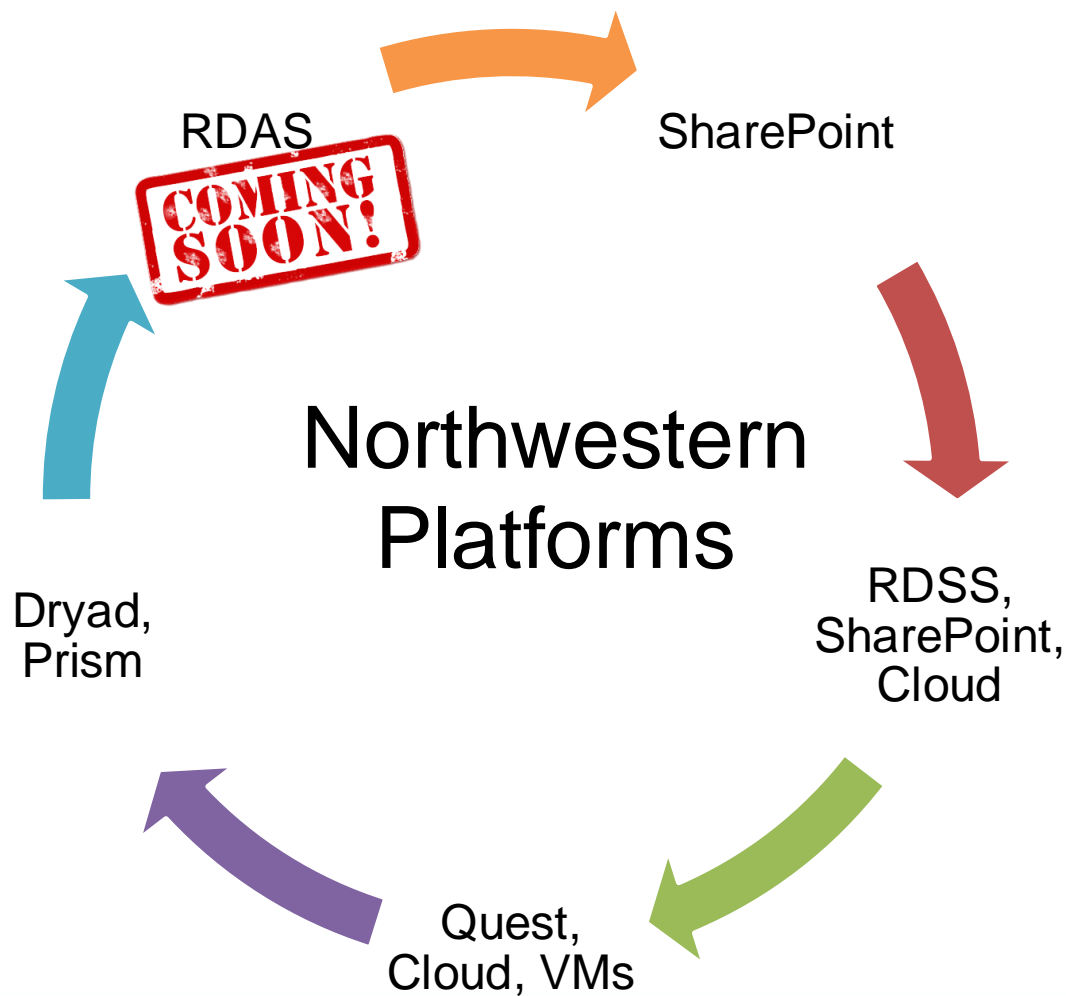The best storage option depends on what you want to do with your data

tl;dr

The best storage option depends on what you want to do with your data
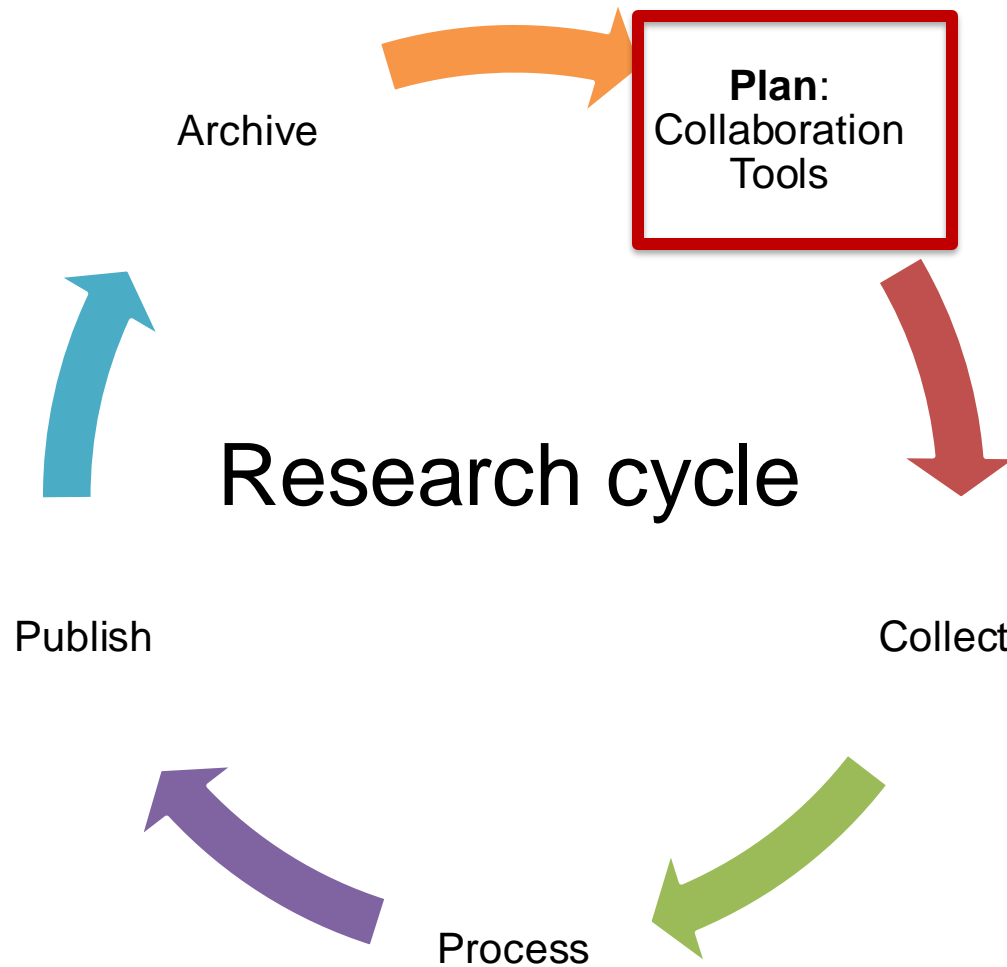
The best option may vary throughout your project.

Research cycle

Archive · Plan · Collect · Process · Publish

Storage Types

Cold storage · Collaboration tools · Hot storage · Fast storage · Data repositories

RDAS

SharePoint

COMING SOON!

Northwestern Platforms

RDSS, SharePoint, Cloud

Dryad, Prism

Quest, Cloud, VMs

Types of storage

Research cycle

Plan: Collaboration Tools

Archive

Publish

Collect

Process

Collaboration Tools

Sharing and collaborating with colleagues

- Cloud-based

- Accessible from anywhere

- Easy to share on and off-campus

- Shared storage for team collaboration
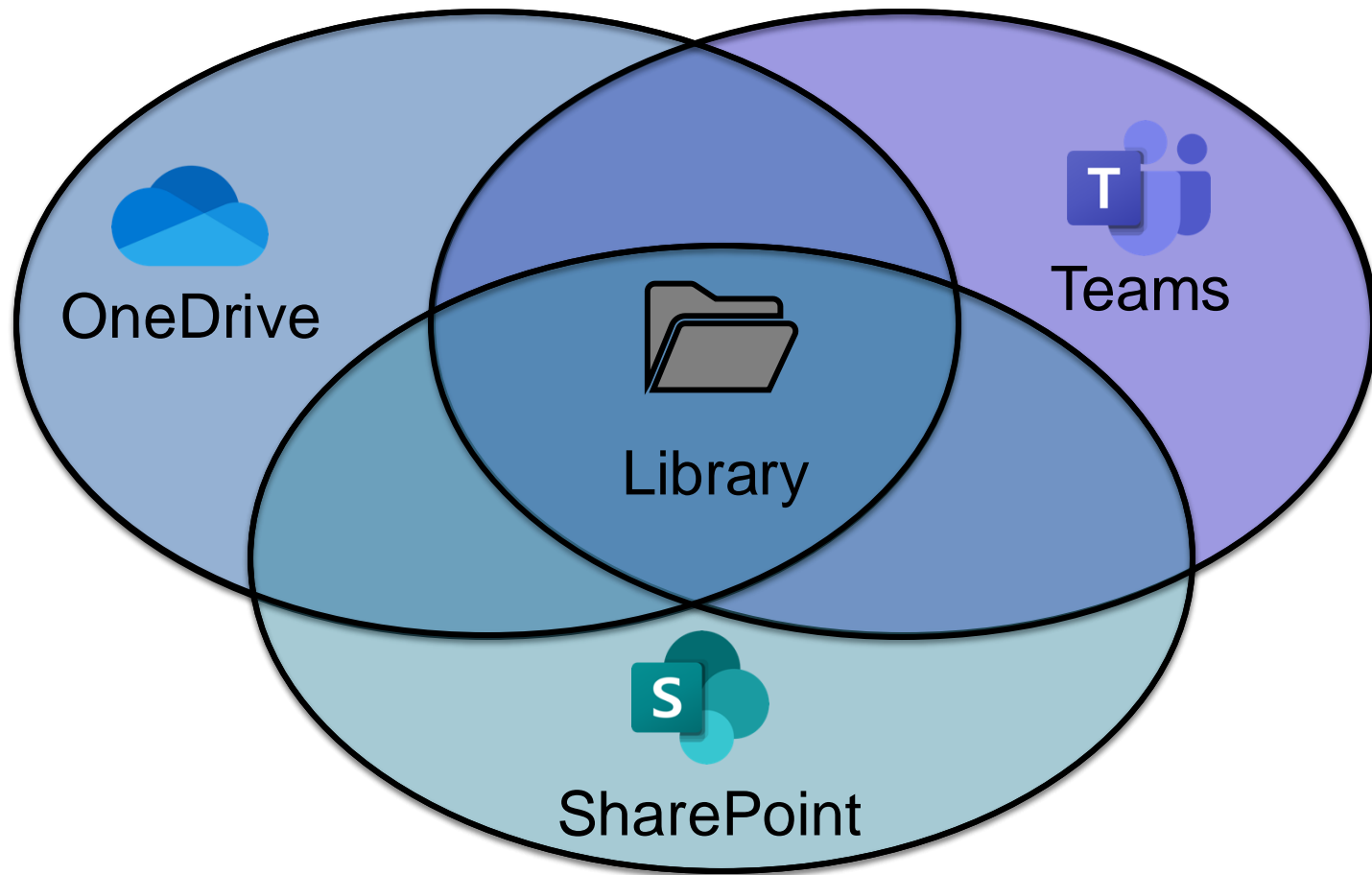
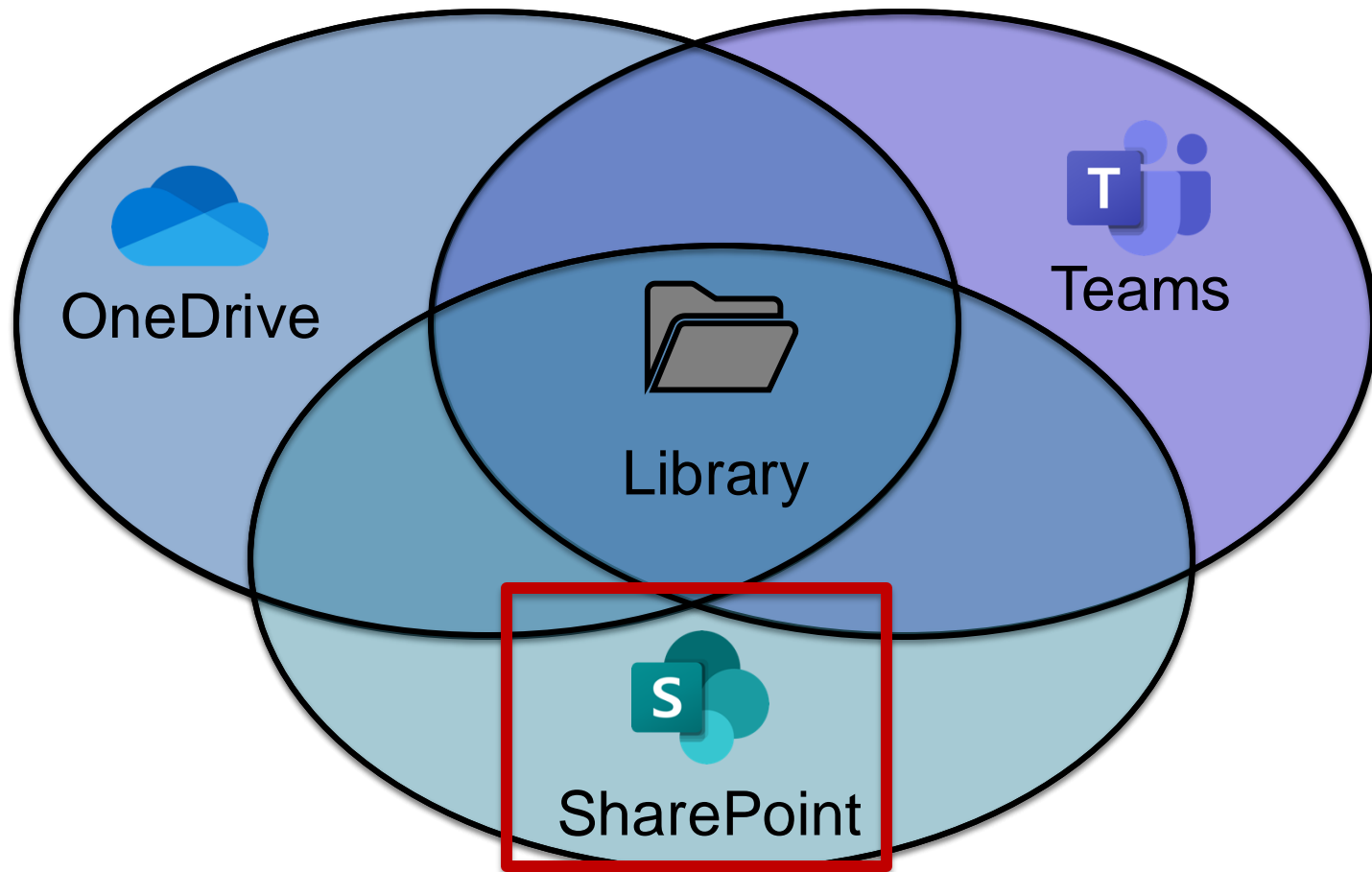# Northwestern uses Microsoft

OneDrive
Personal Files

SharePoint
Team files

Teams
Chat + file storage

SharePoint

Designed for working in teams

- Default access granted at the site level
- Share files/folders with anyone who has a Microsoft account
- Sites include:
  - Pages: Web content
  - Libraries: Store documents
- Security: encryption and auditing
- Document version control
- when you leave the university
  - SharePoint data stays
 - OneDrive data is deleted

SharePoint

SharePoint

When to use it

- Store files that need to be accessible to groups of users

- Collaborate on documents

- Share with people who don't have NetIDs (External collaborators)

When not to use it

- Storing private files (Use OneDrive)

- Long term storage of large data

- Very large files (>250 GB)

- Many small files (>50,000)

Research cycle

Archive

Plan

Publish

Collect:
Active storage

Process
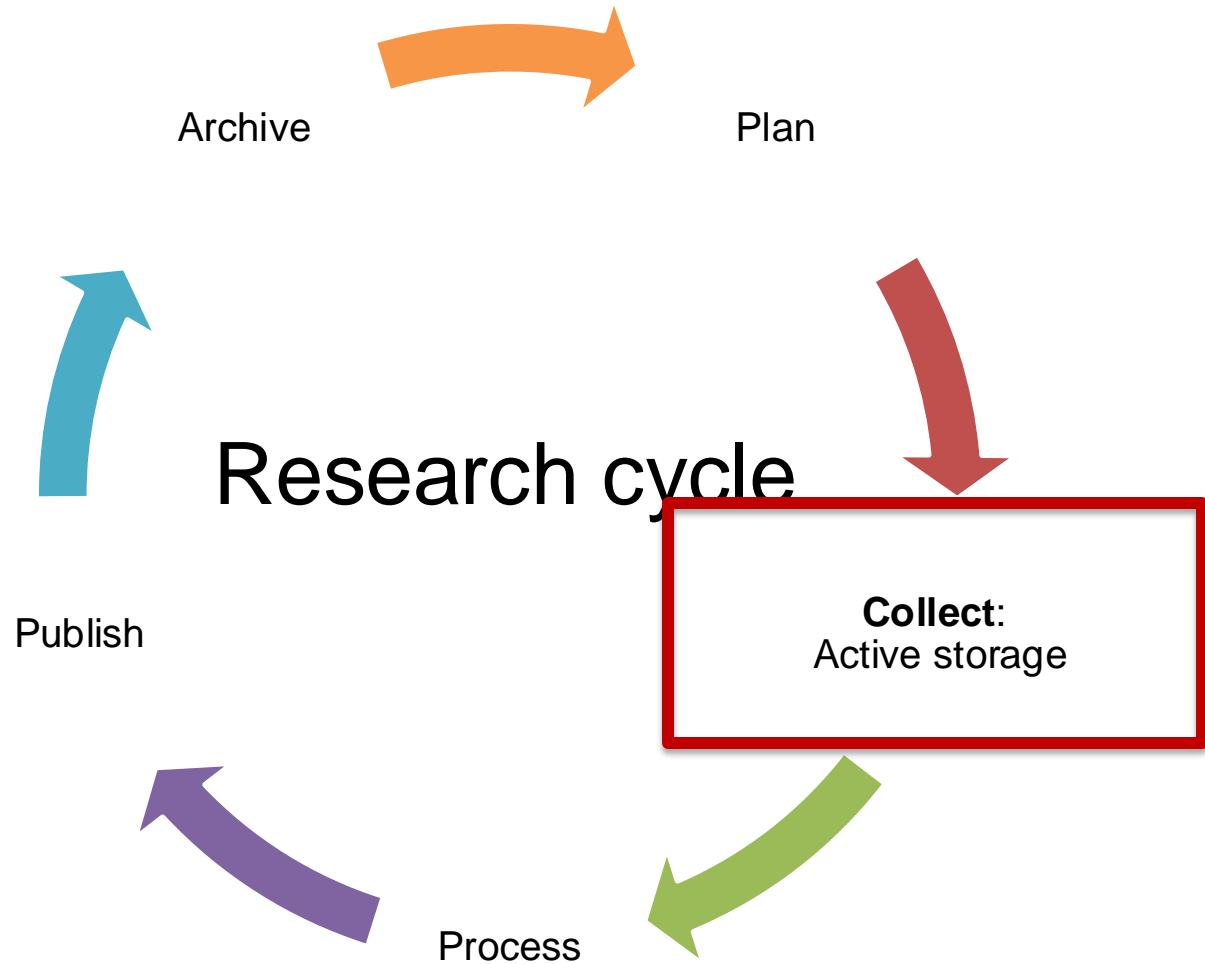
Hot Storage

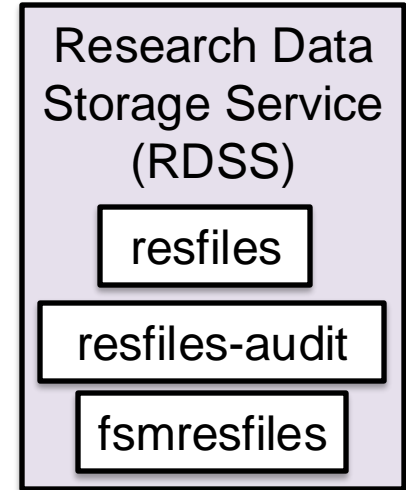For frequently accessed, changing data

- Instant access

- Access for core group members

- Ensure data integrity

- Prevent data loss

Research data storage service (RDSS)

Dedicated research data storage

- Isilon cluster with versioning and off-site backup
- 3 zones
  - **resfiles**: unpublished research data
  - **resfiles-audit**: audited data
  - **fsmresfiles**: Feinberg only
- Map as network drive (on campus/VPN)
- Researchers manage access directly (except fsmresfiles)

https://services.northwestern.edu/TDClient/30/Portal/Requests/ServiceDet?ID=96

Research Data Storage Service (RDSS)

resfiles

resfiles-audit

fsmresfiles

Research data storage service (RDSS)

When to use it

- Access files for active research projects

- Collaborating with a team

When not to use it

- Storing private files

- If you have an unstable internet connection

- Archiving very large data

Research cycle

Archive

Plan

Collect
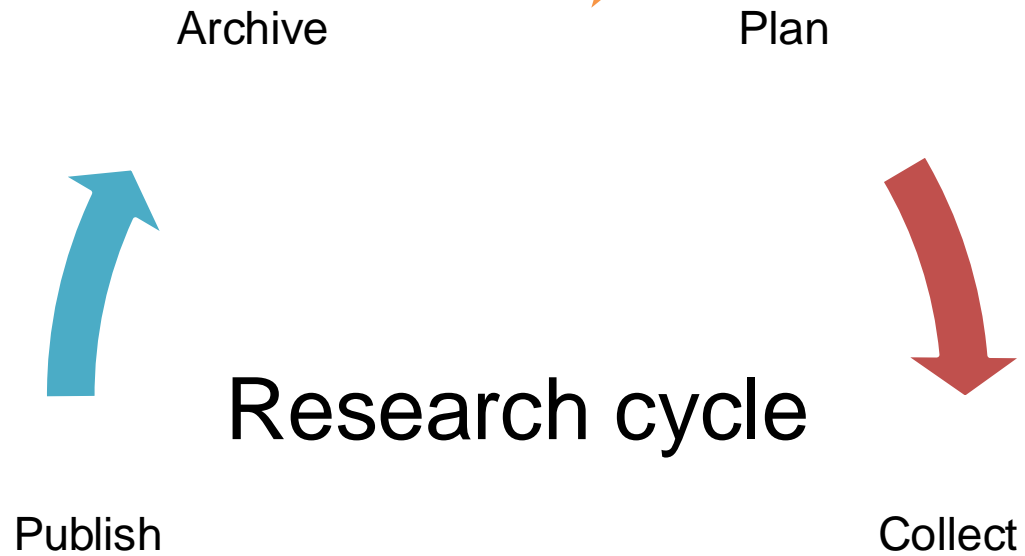
Process: High-performance storage

Publish

Fast Storage

**"Fast" is relative to compute source**

- Computer hard drive (SSD vs. Spinning disc)

- RDSS: Combination of SSD + SD
  - Limited by network speed

- High performance computing
  - High speed network
  - Many nodes that connect to one file system

High Performance Storage

High-throughput data processing

- **Parallel file system**: Storage attached to high performance computing systems

- Fast I/O speeds and network

- Ex: climate modeling, genomic sequencing, [machine learning] and [artificial intelligence], seismic processing
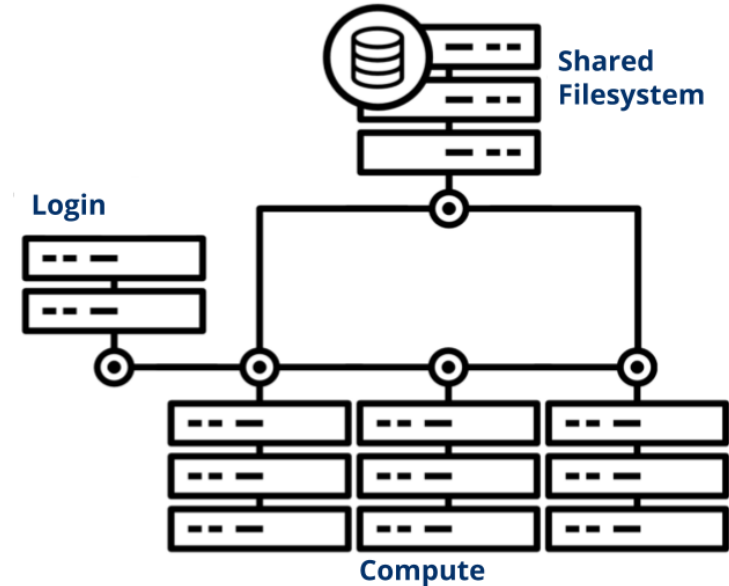
Quest Storage

For processing data on Quest

- Accessible from Quest HPC
- NOT BACKED UP
- Directories
  - **Home** – for individual use, 28 daily snapshots
  - **Project** – for collaboration, not backed up
  - **Scratch** – for short term storage of large, temporary files, not backed up, automatically deleted after 30 days
  - [KB with more details](#)
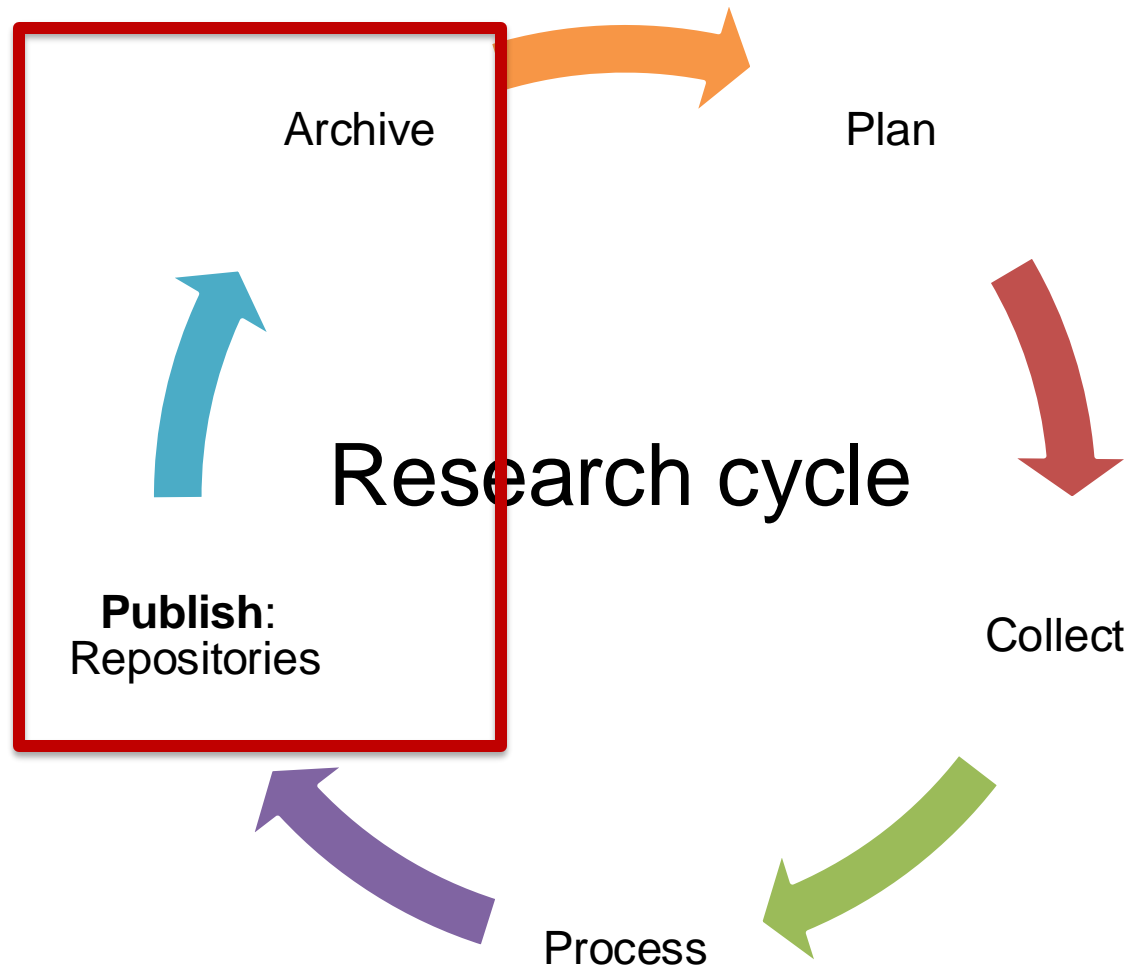
Quest Storage

\home

\project

\scratch
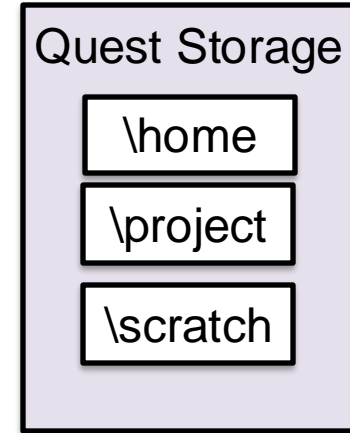
Quest Storage

When to use it

- For any data to be analyzed on Quest

- For data produced on Quest through models

When not to use it

- Any data unrelated to a Quest analysis

- Long term storage of data analyzed or produced on Quest

Research cycle

Archive

Plan

Collect

Process

Publish: Repositories

Public Data Repositories

Sharing data publicly

- Comply with open data requirements

- Provides unique identifiers, searchable metadata, and a license

- Pay at submission, stays indefinitely

Dryad

Open data sharing platform

- Northwestern researchers can share research data repositories of any size at no cost
- Publicly share research data, code and other research materials in one submission process.
- Dryad's curation service checks data integrity
- Dryad's preservation policy guarantees access indefinitely

Dryad

When to use it

- For data you want to share openly on the internet

- When you want your data to be citable

When not to use it

- For data containing personally identifiable human subject information.

- Individual files should not exceed 10 GB

- If you need a license other than CC-0

Archive: Cold Storage

Plan

Research cycle

Collect

Process

Publish
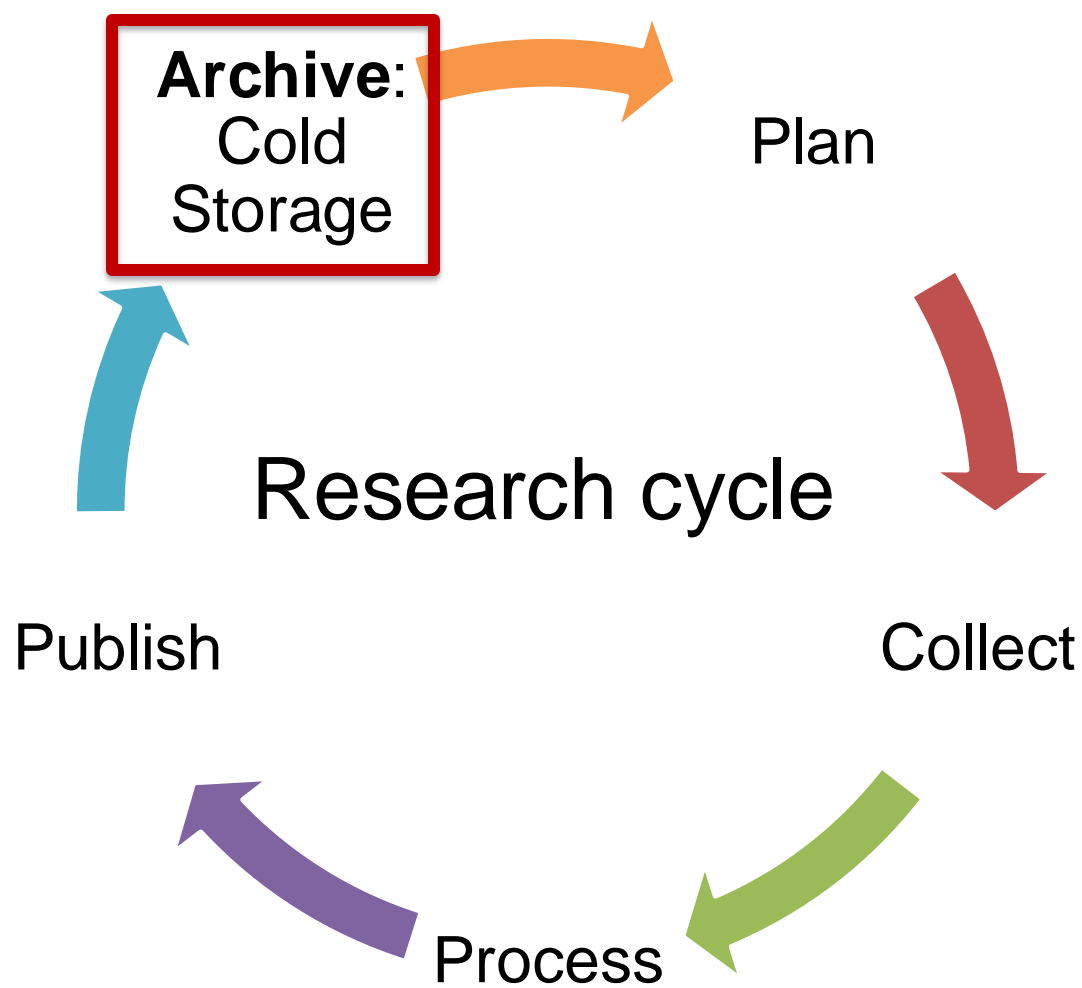
Cold Storage

Retain infrequently used data for long periods of time

- Less expensive storage for infrequently accessed data

- Cost effective way to comply with data retention policies

- Ex: "cold" cloud tiers or tape archives

Tier	Access	Cost/TB/year	Min days
Hot	Frequent (multiple times a day) Instant retrieval	~$240	none
Cool	Infrequently (weekly or monthly) quick retrieval	~$120	~30
Cold	Rarely (1-2 times a year) slower retrieval	~$48	~90
Archive	Very rarely (<1x per year) slowest retrieval	~$12	180-365

Northwestern data retention policies

Data Type	Retention Period
All Northwestern research data	At least three years
Data generated by students	Until the student graduates or leaves Northwestern and all papers are published
Data supporting patent applications	Until the patent process is complete
Data subject to litigation or audit	Until the situation is resolved
Data subject to HIPAA or under a HIPAA waiver	Six years past the end of project completion

See Northwestern's research data policy, Northwestern's Retention of University Records policy and its associated records retention schedule, the IRB investigator manual, and the University Patent and Invention policy for more information.

Research Data Archival Service

Retain infrequently used data

- Cloud-based (Amazon S3)
- Submit data to RCDS
- Specify retention period
- RCDS staff handles administrative work (cloud accounts, billing)
- Request retrieval
- Decide what to do at end of retention period

COMING SOON!

Research Data Archival Service

COMING SOON!

When to use it

- Store data that is accessed <1x per year

- When you need to control storage costs

- Outsource administrative tasks re: cloud storage

When not to use it

- For frequently accessed data

- If you want direct access to retrieve your data

Poll
Do you vary where you store your data throughout the project?

How to Choose
(Questions to ask yourself)

Questions to ask yourself

- **How much** data do you have?
- What is your **budget**?
- Who needs **access**?
- What **regulations** apply to your data?
- How will you **protect** your data?
- What does your **workflow** look like?

How much data do you have?

Estimate how much data you have and/or will produce

- Total capacity
- Number of files
- Largest individual tile

	Capacity	Individual file size	Number of Files
Quest	1 – 2 TB free, buy more	9 exabytes	9 quintillion
RDSS	Pay per TB	16 TB	billions
Share Point	25 TB	250 GB	30,000,000/site
Cloud	Pay per use	~ 5 TB	No limit

What's your budget?

- Can budget for data storage costs in grant applications

- Talk to the project PI about budgeting for storage

- Price can reflect performance and other features (eg: backups, versioning)

	Free Tier	Cost
Quest	1 – 2 TB	$195/TB/5 years
RDSS	none*	$100/TB/year
SharePoint	25 TB	none
Cloud	none	Pay for what you use

Cost and Capacity

	Quest	RDSS	SharePoint	Public Cloud
Free Tier	1 – 2 TB	none	25 TB	none
Cost	$195/TB/ 5 years	$100/TB/year	none	It depends
Individual file size	9 exabytes	16 TB	250 GB	~ 5 TB
Number of Files	9 quintillion	billions	30,000,000/ site	unlimited

Worksheet:

How much data will you have?
What is your budget?

Who Needs Access?

Access credentials vary between systems

- Can all your collaborators get necessary credentials?

	Credential	Notes
Quest	Quest account	Easy within Northwestern, can request Guest NetID for collaborators
RDSS	NetID	
Share Point	Microsoft account	Many universities use MS, anyone can get one
Public Cloud	IAM roles	You create access credentials

Who Needs Access?

Granularity of permissions vary between systems

- **Default Access –** who gets access to what by default

- **Sharing –** how to share beyond the default group

- **Roles –** What you can do in the system

	Default Access	Sharing	Roles
Quest	Home – you Project – team Scratch - you	Linux permissions	Read Write execute
RDSS	Whole share	Not available	Read write Read only
Share Point	Whole Library	By file/folder	View edit review
Public Cloud	Configurable	Configurable	Configurable

Access

	Quest	RDSS	SharePoint	Public Cloud
Credentials	Quest (Linux) account	NetID	Microsoft account	IAM users
Default access	Whole project	Whole Share	Whole Library, By file/folder	Configurable
Sharing	Grant access to files/folders Globus	none	By file or folder	Configurable
Roles	read/write/execute	Read/Write Read only	View, edit, review	Configurable

Worksheet:

Who needs access?

What regulations apply to your data?

Depends on data type and where you store it

- Northwestern Policies
- Storage system policies
- Data use agreements (DUAs)
- Federal and state regulations

Basic Safeguards

Northwestern's minimum requirements for all data

- **Basic safeguards**: 17 protective measures to keep your data confidential, accessible and intact

- Northwestern-approved systems satisfy these basic safeguards

- If build your own systems (eg: cloud), you should implement the safeguards yourself
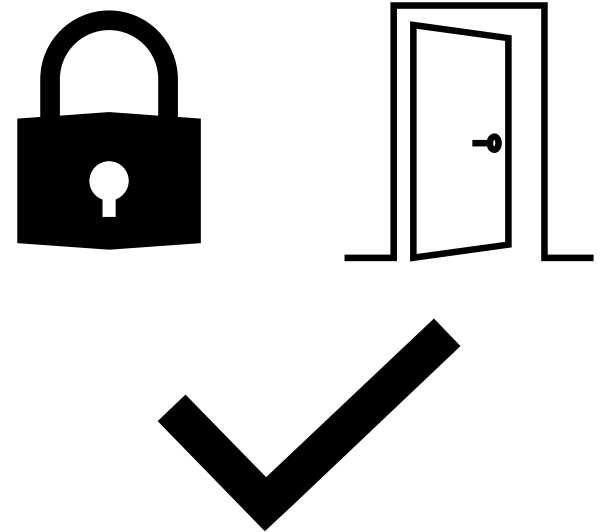
Things you can do

Use Northwestern-supported data storage systems

Use servers, computers, and mobile devices that require passwords or PINs

Use individual accounts and don't share your password

Grant users access rights appropriately: Be careful what you share

Use secured and trusted networks or a VPN

Update your OS and use antivirus software with ransomware protection.

System policies

Every system has its own policies

- Quest's data policy disallows Personally Identifiable Information (PII)

- RDSS restricts certain types of data to certain zones
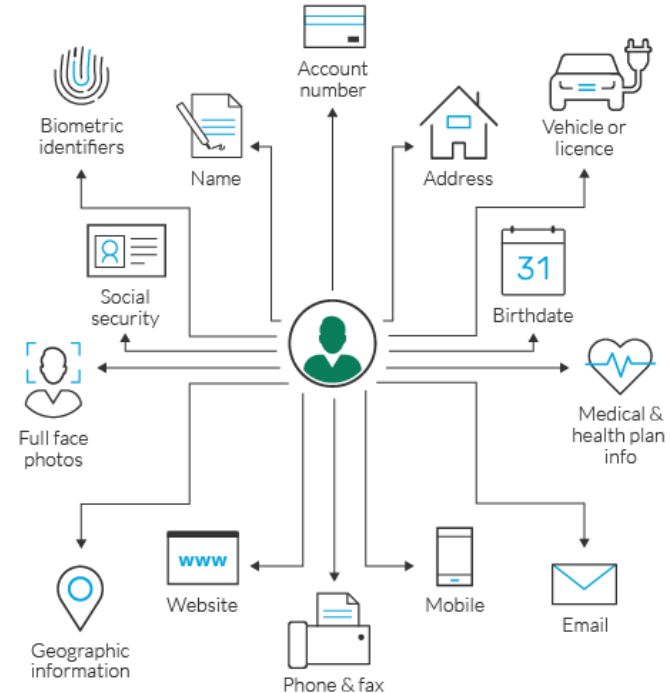
Data use agreements

DUAs are contracts that have specific terms

- When data comes from a third party, a DUA specifies what you can and cannot do with the data

- **Specific**: compliance with a named security framework, encryption, access control

- **Vague**: "Must be stored on a university run system"

Legal requirements

Federal and state laws place restrictions on certain types of data

- Only certain Northwestern systems can accommodate these restrictions

- Eg: protected health information (PHI), controlled unclassified information (CUI), Biometric data (in Illinois)

- Highly regulated data like government classified data and export-controlled data cannot be kept on Northwestern storage systems

Legal requirements

Federal and state laws place restrictions on certain types of data

- Only certain Northwestern systems can accommodate these restrictions

- Eg: protected health information (PHI), controlled unclassified information (CUI), Biometric data (in Illinois)

- **Highly regulated data like government classified data and export-controlled data cannot be kept on Northwestern storage systems**

Ask the project PI if there are any special requirements for handling your research data.

Contact the Information Security Office if you're not sure how to meet your specific requirements.

What **regulations** apply to your data?

How will you protect your data?

Threats to your data

- Hardware failure (bit rot)
- Human error (accidental deletion)
- Natural Disaster (my building flooded)
- Cybersecurity threat (ransomware)
- Software glitch (my code overwrote my raw data)

Data durability features

Feature	Meaning	Protects against
Redundancy	Keeping multiple copies in multiple locations	Hardware failure, human error, natural disaster, cybersecurity threat
Versioning	Keeping multiple versions of files	Human error, cybersecurity threat
Restoring files	Retrieving a file after accidental deletion or modification	Human error
Data integrity checks	Using an algorithm to check for file degradation	Bit rot, Hardware failure

Data durability features

	Quest	RDSS	SharePoint	Public Cloud
# Copies	1	2	3	2-3
Replication	no	2 locations	2 regions	configurable
Version frequency	no	Daily	"every few minutes"	configurable
# Versions	no	28	Max 50,000 (can set by owner)	configurable
Restore deleted items	no	Via snapshot	Recycle bin for 93 days, 14-day grace period	No
Data integrity checks	no	Continuous	Every 14 days	Regularly

Worksheet:

How will you protect your data?

What does your workflow look like?

Use storage that connects with where you do your work

- Where is your data produced

- Where will you analyze the data?

- How fast does the connection need to be?

- Where will it be archived?
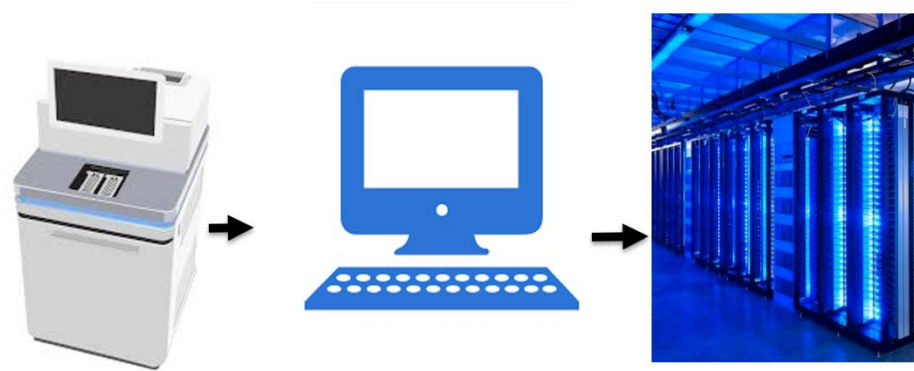
Direct connection to Compute

Compute Source	Where to store data	Comments
Your computer	Your hard drive	Might not be big enough Needs a backup
	External drive	Might not be fast enough
	Network Drive (RDSS)	Might not be fast enough
	SharePoint	Hard to get big data in/out
Quest	Quest storage	Only option Needs a backup location
"The Cloud"	Cloud Storage	Pay per use

Data transfer: Globus

- Designed for large file transfer
- Transfer and sync data between "collections"
  - Many NU storage systems
  - Your computer
  - More to come this fall
- Error handling

```
Amazon S3  ←→  OneDrive-SharePoint

Quest  ←→  Your computer

resfiles | fsmresfiles | resfiles-audit
RDSS
```
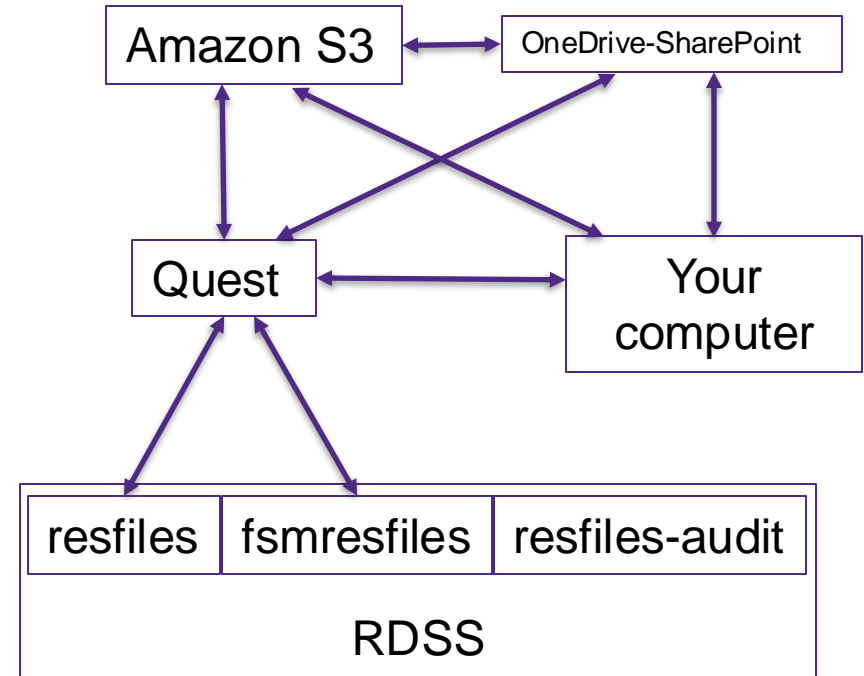
Example Workflow

Instrument

Quest Storage

SharePoint

?
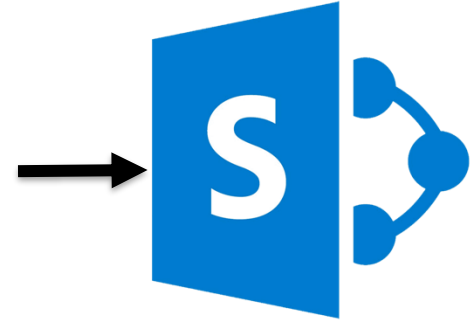
You're using data produced at a core facility.

You need to analyze it on Quest, then transfer the results to SharePoint to collaborate on a paper

Direct to Quest

Instrument

Quest Storage

SharePoint

Direct to Quest

Instrument
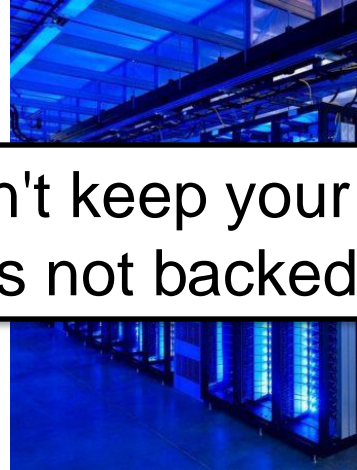
Quest Storage
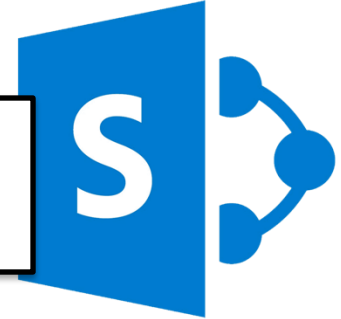
SharePoint

- Core facility won't keep your data
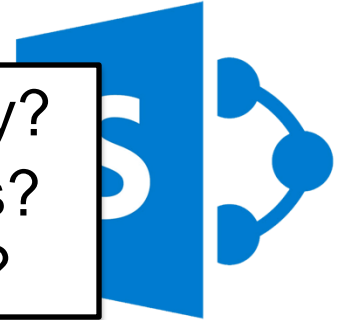- Quest Storage is not backed up

Via your computer

Instrument → Your computer → Quest Storage → SharePoint
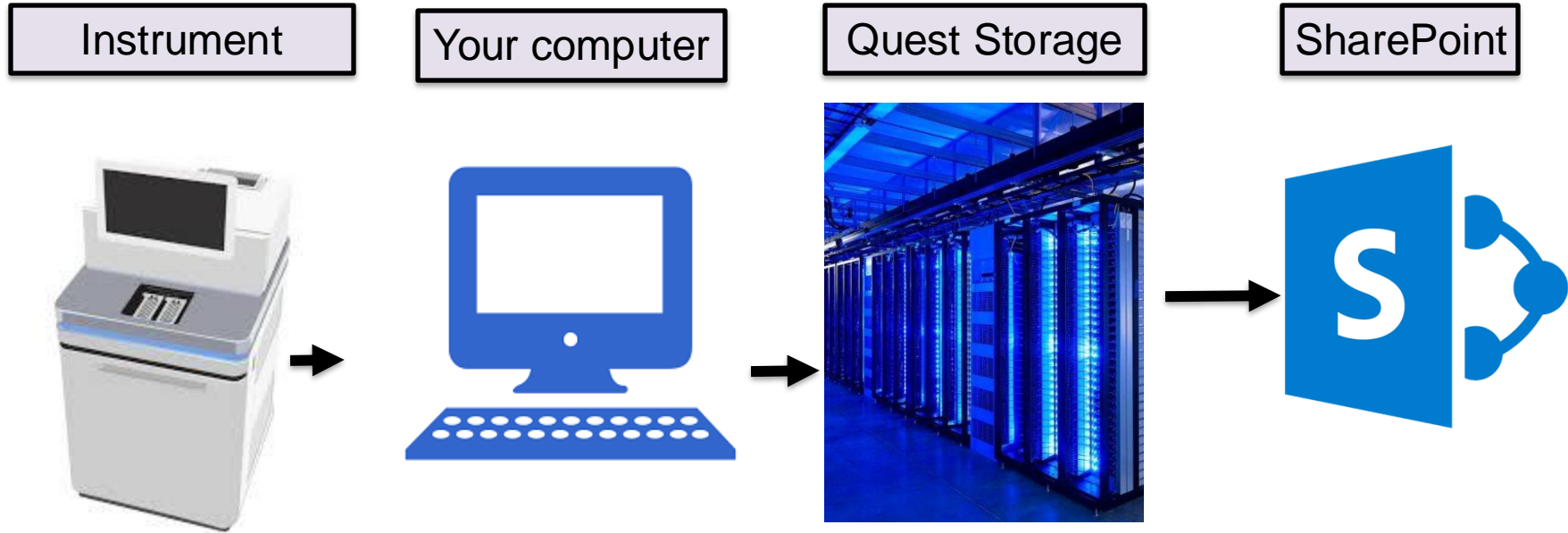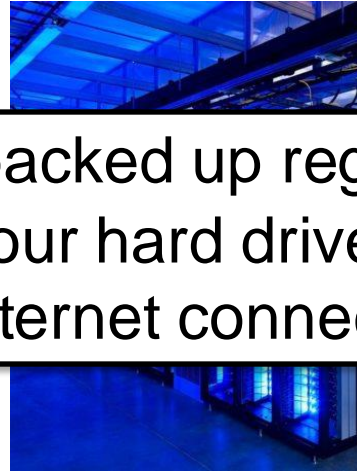
Via your computer

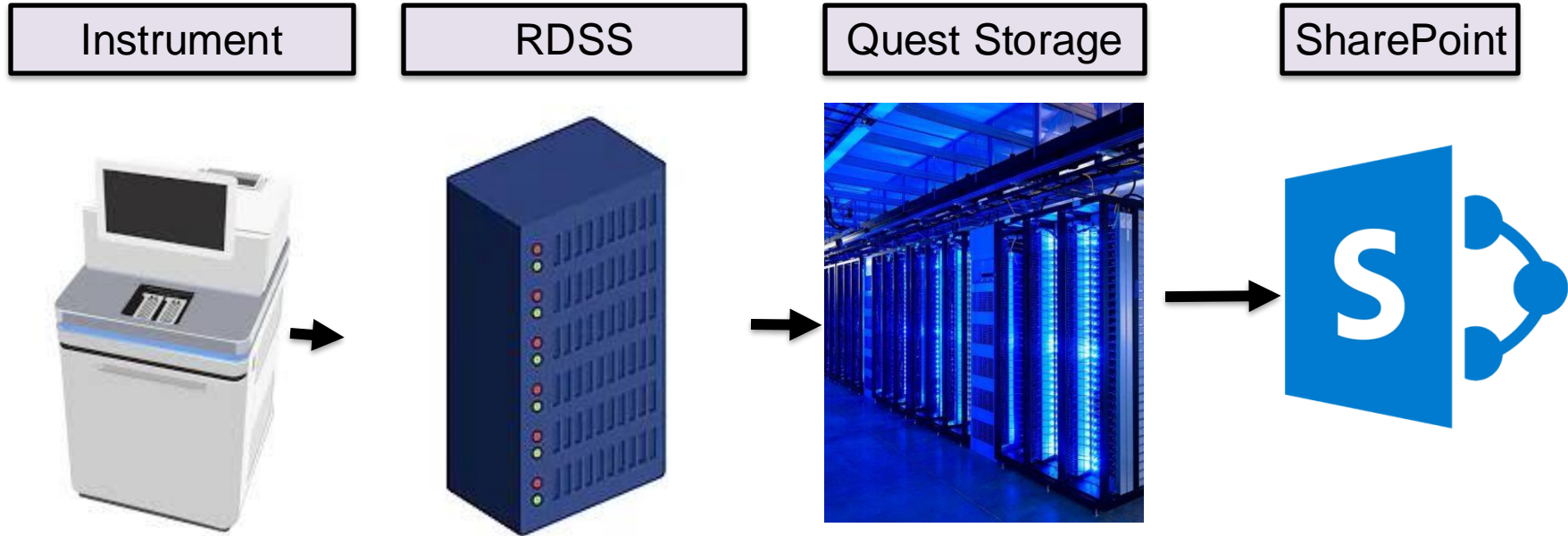Instrument

Your computer

Quest Storage

SharePoint

- Is your computer backed up regularly?
- What happens if your hard drive dies?
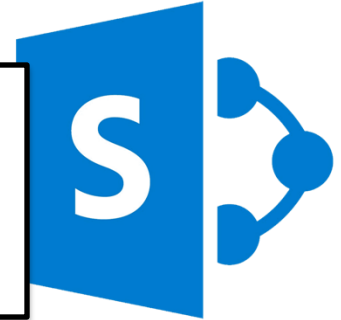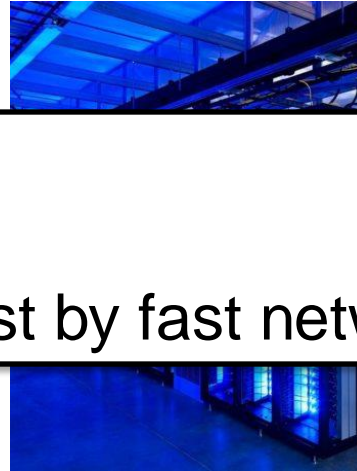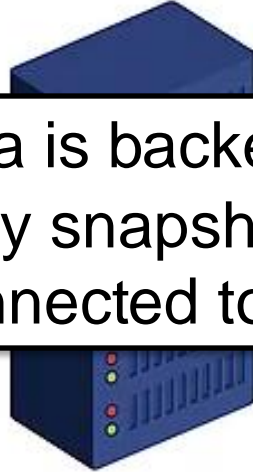- How fast is your internet connection?

Via RDSS/FSMResfiles

Instrument → RDSS → Quest Storage → SharePoint

Via RDSS/FSMResfiles

Instrument   RDSS   Quest Storage   SharePoint

- Data is backed up
- Daily snapshots
- Connected to Quest by fast network

Worksheet:

Draw your workflow

Reach out!

FIND WHAT YOU NEED

PLANNING
- Writing a Data Management Plan
- Protecting the Sensitive Information in My Data

DATA COLLECTION AND STORAGE
- Choosing Appropriate Storage
- Documenting Your Research
- Transferring Data to or from Northwestern
- Sharing Data with an External Collaborator

DATA SHARING AND ARCHIVING
- Making Your Data Reusable
- Sharing Data Publicly
- Archiving Data When a Project is Done

SUPPORT AND RESOURCES
- Talk to a Data Management Expert
- Northwestern Research Data Management Resources
- External Research Data Management Resources

Research Data Management Website

Office Hours: Every Monday
3 p.m. – 4 p.m.
Mudd Library,
GIS Lab

Emails:
researchdata@northwestern.edu

Consultation Calls

Resources

Email: researchdata@northwestern.edu to get help on any data management topic

Request resources

- Request RDSS storage
- Request SharePoint site
- Request Quest Allocation
- Request Public Cloud accounts
- Submit data to Dryad
- Use an Electronic Research Notebook

Useful links

- Northwestern Research Data Management Website
- Information Security: Protect your research

Questions?