February 14, 2024 | 10:30am-11:30am
Virtual Workshop

Storage Scores: Store & Back Up Data at McMaster

u.mcmaster.ca/scds-events

RDM

SCDS

Library

McMaster University

Storage Scores: Store & Back Up Data at McMaster

John Fink, MLS
Digital Scholarship Librarian
jfink@mcmaster.ca

Danica Evering, MA
Research Data Management Specialist
rdm@mcmaster.ca

Research Data Management Workshop Series
February 14, 2024

Lewis & Ruth Sherman Centre for Digital Scholarship
scds.ca

McMaster University Library

McMaster University is located on the traditional Territories of the Mississauga and Haudenosaunee Nations, and within the lands protected by the “Dish With One Spoon” wampum agreement.

Laslovarga, “Webster Falls in Winter, Waterdown, Hamilton, Ontario, Canada - Spencer Gorge / Webster's Falls Conservation Area,” 23 January 2011, Wikimedia Commons - https://commons.wikimedia.org/wiki/File:Waterdawn_Webster_Falls_in_Winter8.jpg

Code of Conduct

The Sherman Centre and the McMaster University Library are committed to fostering a supportive and inclusive environment for its presenters and participants.

As a participant in this session, you agree to support and help cultivate an experience that is collaborative, respectful, and inclusive, as well as free of harassment, discrimination, and oppression. We reserve the right to remove participants who exhibit harassing, malicious, or persistently disruptive behaviour.

Please refer to our code of conduct webpage for more information: scds.ca/events/code-of-conduct/

Session Recording and Privacy

This session is being recorded with the intention of being shared publicly via the web for future audiences. In respect of your privacy, participant lists will not be shared outside of this session, nor will question or chat transcripts.

Questions asked via the chat box will be read by the facilitator without identifying you. Note that you may be identifiable when asking a question during the session in an audio or visual format.

Certificate Program

The Sherman Centre offers a Certificate of Attendance that rewards synchronous participation at 7 workshops. We also offer concentrations in Data Analysis and Visualization, Digital Scholarship, and Research Data Management.

Learn more about the Certificate Program: https://scds.ca/certificate-program

Verify your participation at a session: https://u.mcmaster.ca/verification

At an unspecified point during the workshop, a code will be read aloud. This is the answer to the third question of the form.

Research Data Management Workshops

Register for upcoming RDM events: https://rdm.mcmaster.ca/events

Mar. 20: “How to Implement Encryption to Protect Your Research Data”

Apr. 17: “Sensitive Data Management”

May 14: “Data Management Plan (DMP) Bootcamp”

Jun. 18: “Data Deposit Bootcamp”

Hello! A bit about us:

Danica Evering, MA (they/them)
My background is in **social practice art, community-based research**, **communications studies**, and **medical laboratory healthcare**.

I have an MA in **Media Studies** from Concordia University.

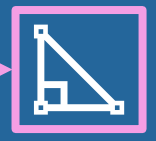
John Fink, MLS (he/they)
Pro at complex and innovative **systems administration and project management**! Also has an interest in the **maker/hacker element** in digital scholarship, frequently spotted tinkering with esoteric hardware.

Outline

Approaches to data storage - matching needs with platforms

McMaster solutions for active data management

- Local storage
- Networked storage
- Cloud storage

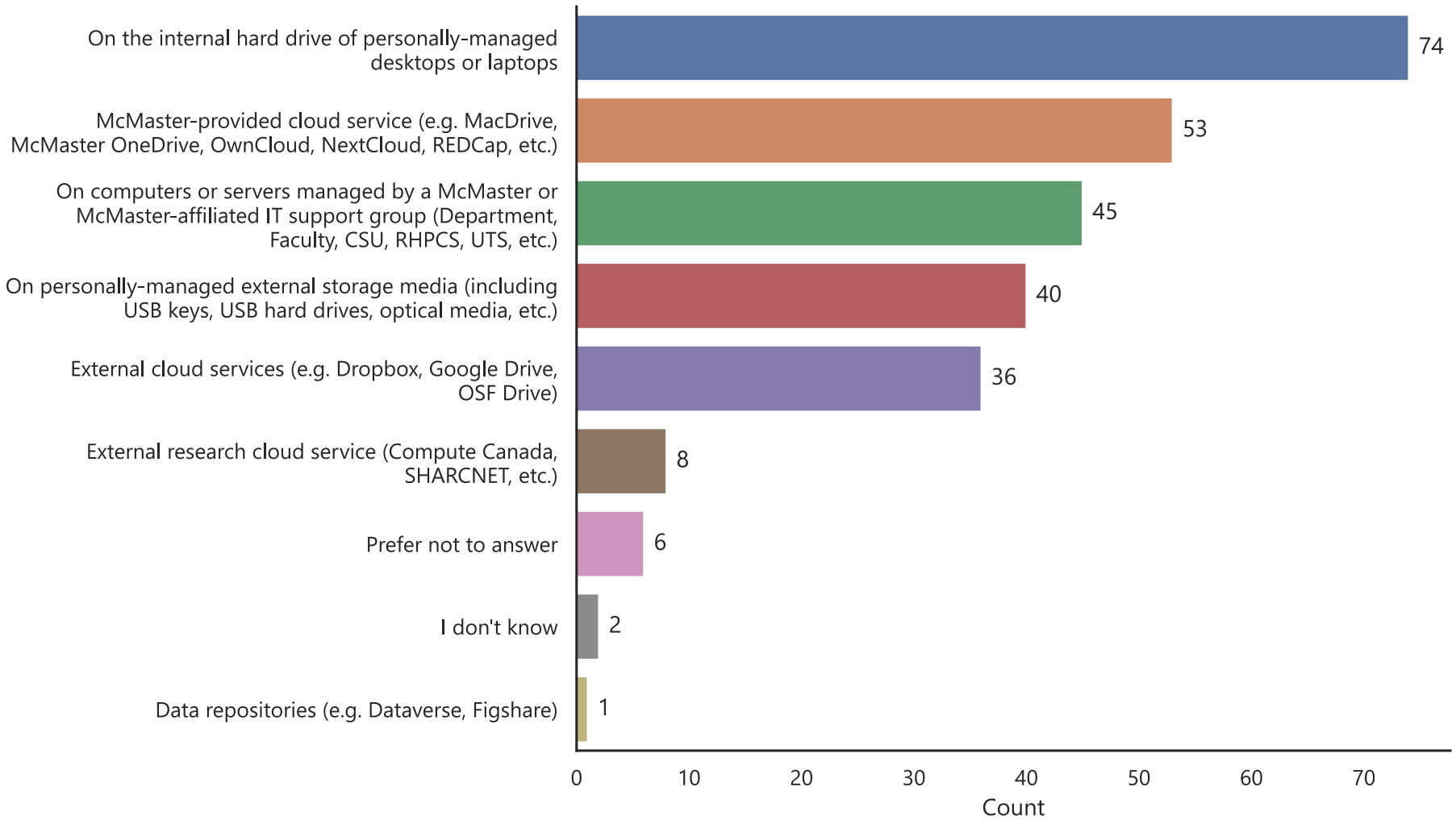


Digital Research Alliance of Canada – storage and computing

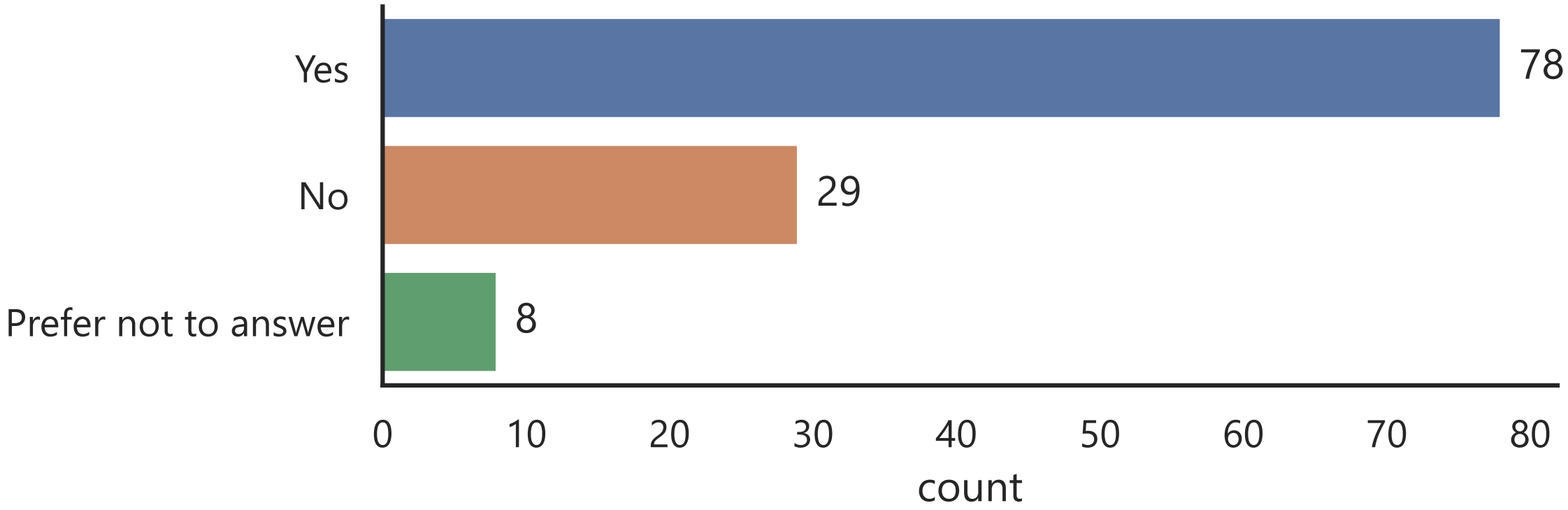
Data security precautions and sensitive data management

Data publishing, deposit, and archiving

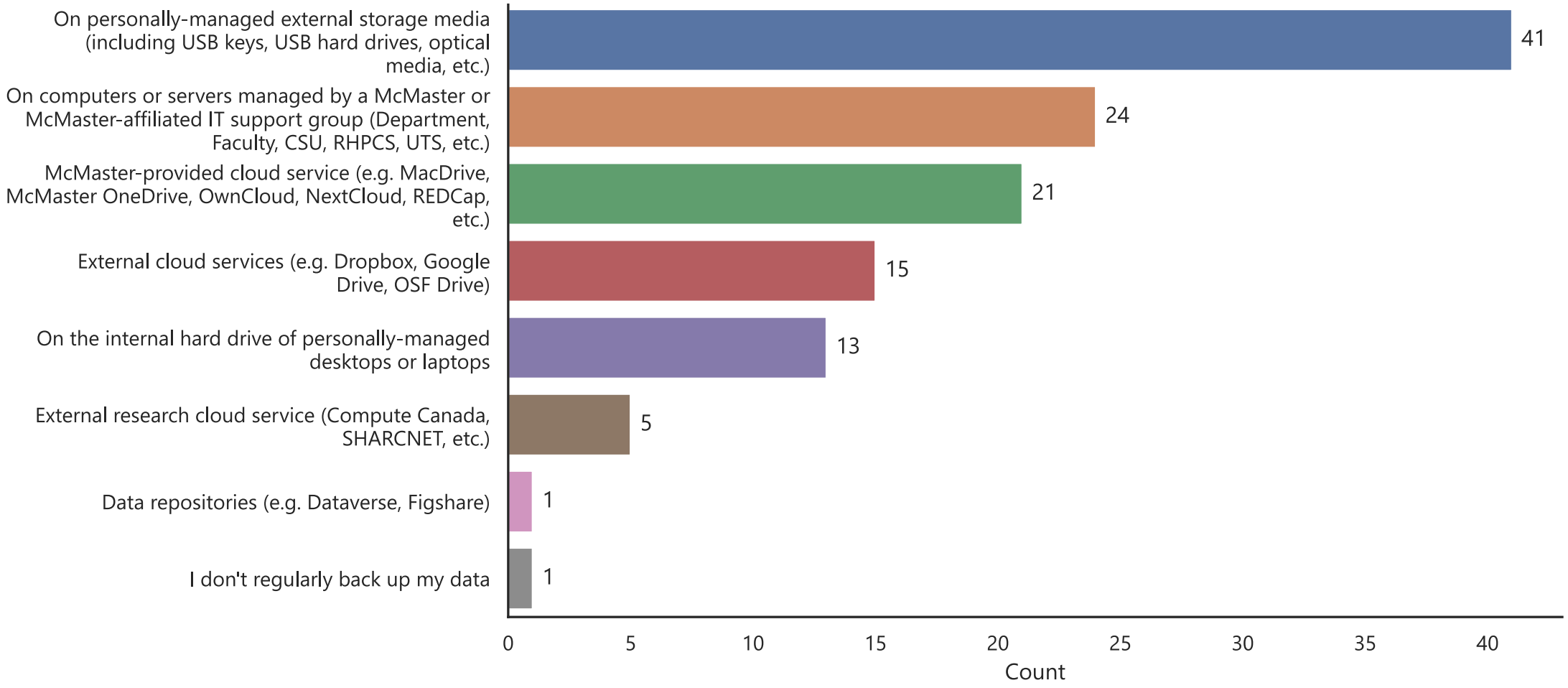
Where do you store your research data?



Are backup copies of your research data maintained separately from your primary storage?



Where is your data backed up?



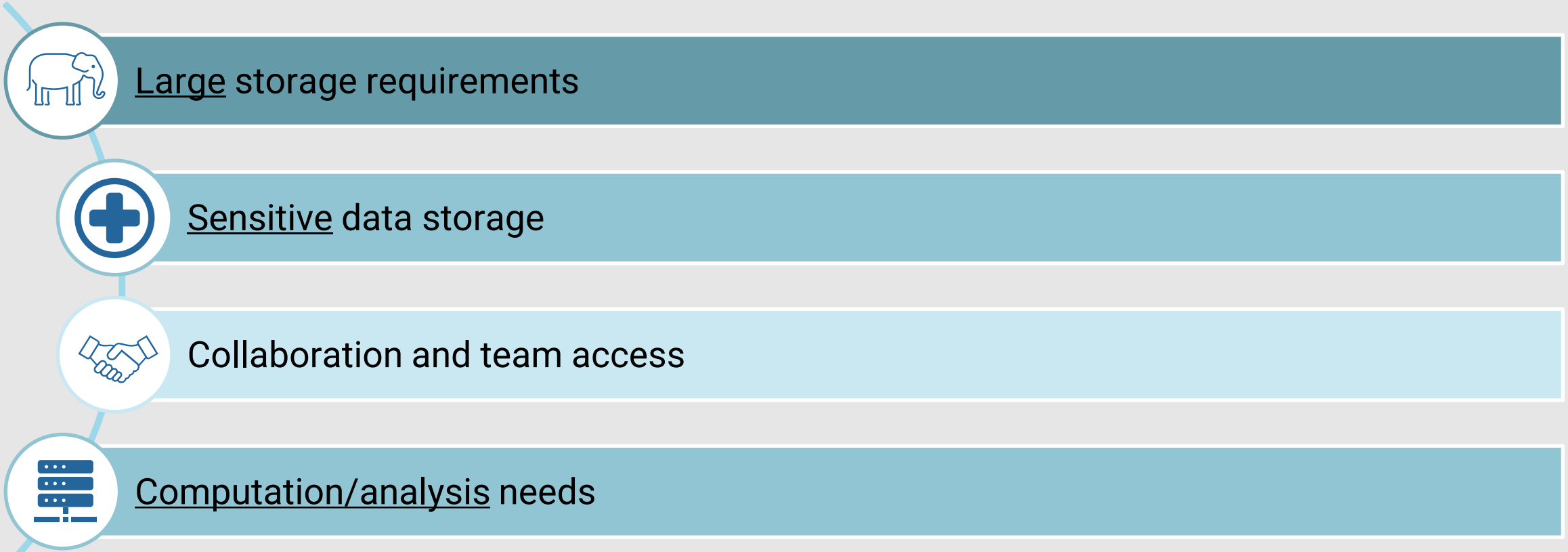
A Data Management Plan (DMP) outlines a storage and back-up plan…for you and your team.

- Anticipate volume, length of time, storage location, collaboration, and back ups.
- Prepare for future stages of research including potential data sharing, deposit, and archiving.
- Research is a team effort – collaborate on your DMP and use it as an onboarding document to avoid students and staff storing their data in separate places.
- https://rdm.mcmaster.ca/plan

McMaster University | Library

Data Storage Principles

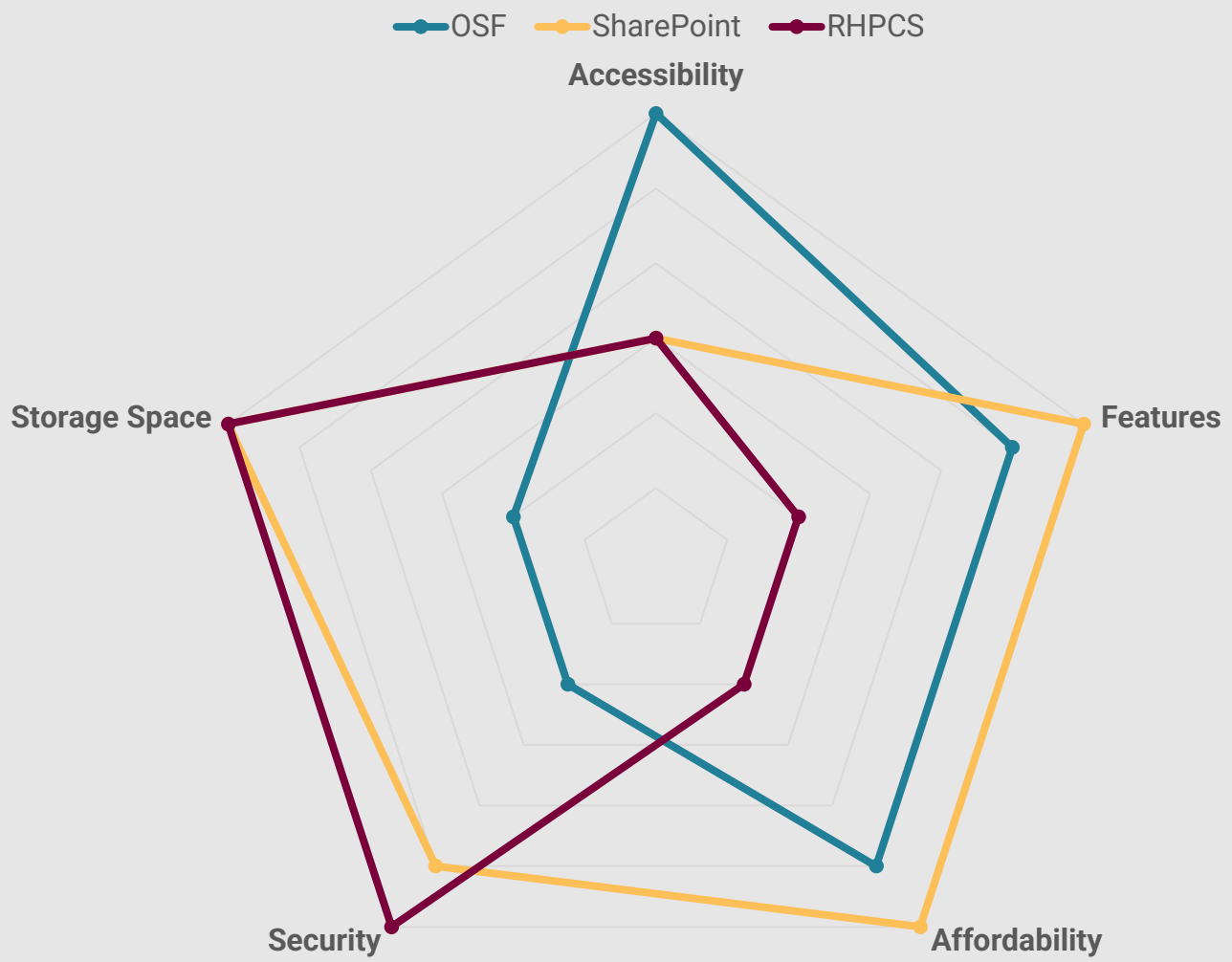
Every dataset has a different set of **specific needs**. This can relate to size, sensitivity, complexity, and more:



Data Storage Variables

Researchers selecting a data storage services balance several different variables: Accessibility, Feature Set, Security, Storage Space, and Affordability.

(Features can include things like computation capabilities or collaboration features)

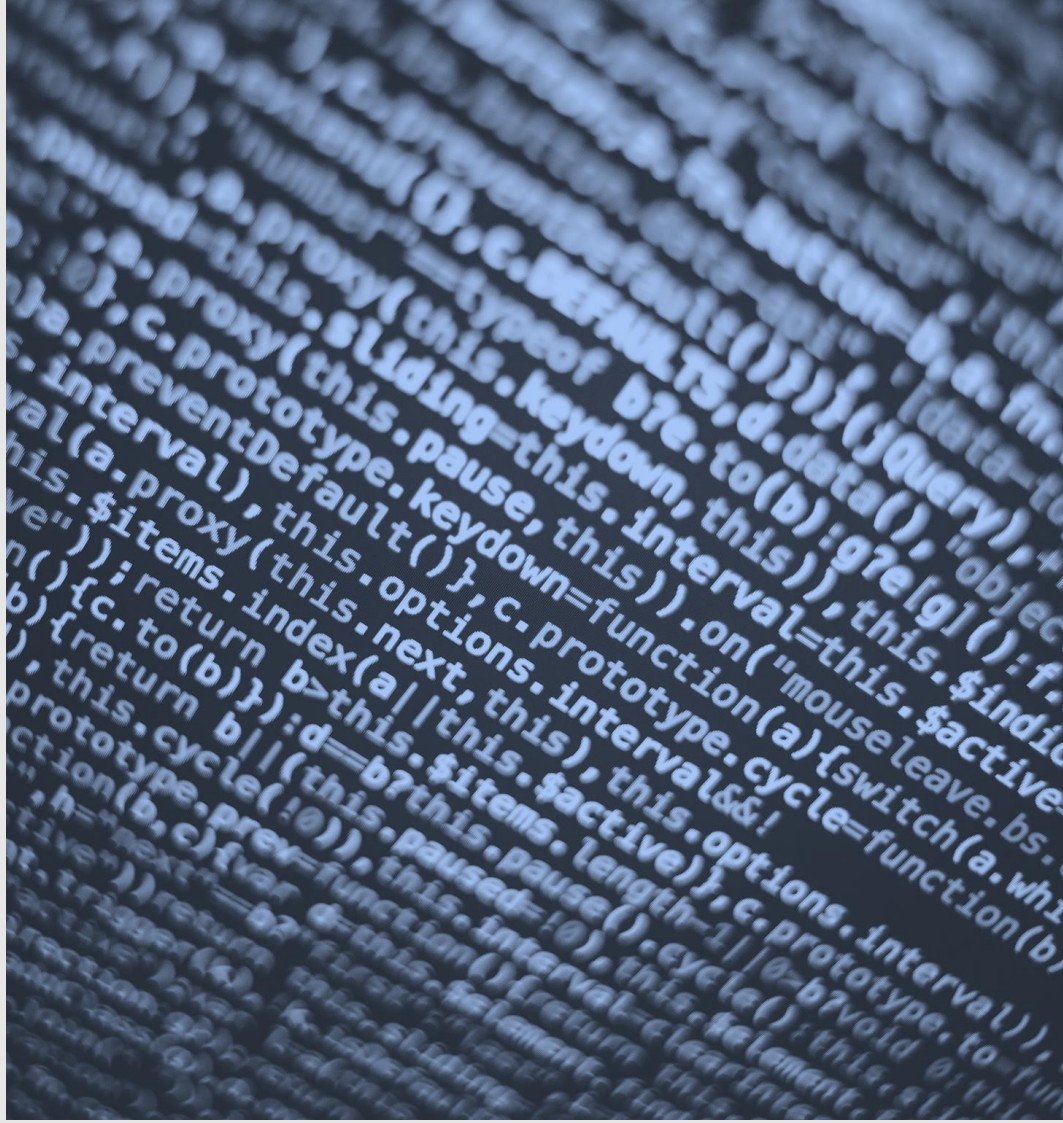


Data Storage Principles

- Although this isn't an absolute rule, the more secure your data is, the less accessible it will be.
- Features **like multi-factor authentication**, **encryption**, and **password** protection all make it somewhat more difficult to access data.
- However, in return for small inconveniences, your data is more secure.
- In thinking about your research project, ask yourself:
  - **How secure does my data need to be?**
  - **What other features am I looking for?**

Backup Strategies – *Case Study: Mount St. Mary College*

- On December 20, 2022, Mount St. Mary College identified and stopped a ransomware attack.
- Russian-speaking hacking group Vice Society took credit for this attack.
- University received a ransom demand; data was later published on the dark web.
- Data are valuable – to the researcher, the school, or to others.
- "Double extortion" attack = data theft + encryption. Security stops data theft, and backups mean you still might have a copy of files that are not encrypted.



"Universities and Colleges Cope Silently with Ransomware Attacks," CSO Online, accessed September 18, 2023, https://www.csoonline.com/article/574739/universities-and-colleges-cope-silently-with-ransomware-attacks.html.

Backup Strategies (3-2-1)

A good data storage plan needs to balance accessibility and convenience against security and reliability.

3 Copies of your data (at least!)

Example:
1 copy stored locally on **hard drive** for analysis
1 copy stored on **cloud storage** platform
1 copy stored in a **secure campus drive**

2 Copies are on-hand (easily accessible) on different storage systems (internal hard drive, external hard drive, USB, etc.)

- a "**production**" (working copy)
- a "**production backup**"

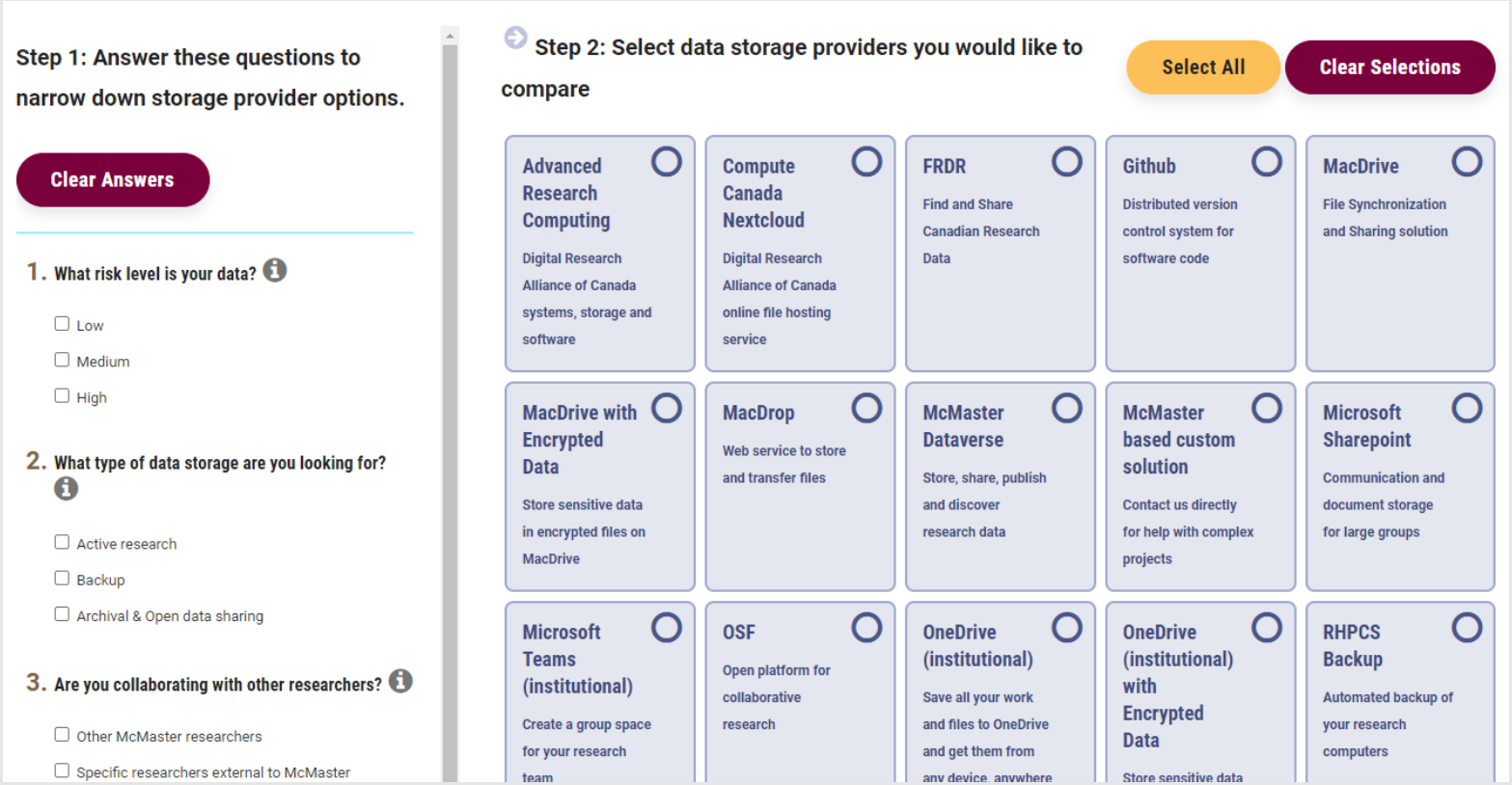
3-2-1-1-0 can provide extra protections from ransomware and errors!
1 – offline or "air-gapped"
0 – backups must contain zero errors

1 Copy is in another location ("off-site") from the others with a ***trusted*** service provider

Research Data Storage Finder Tool

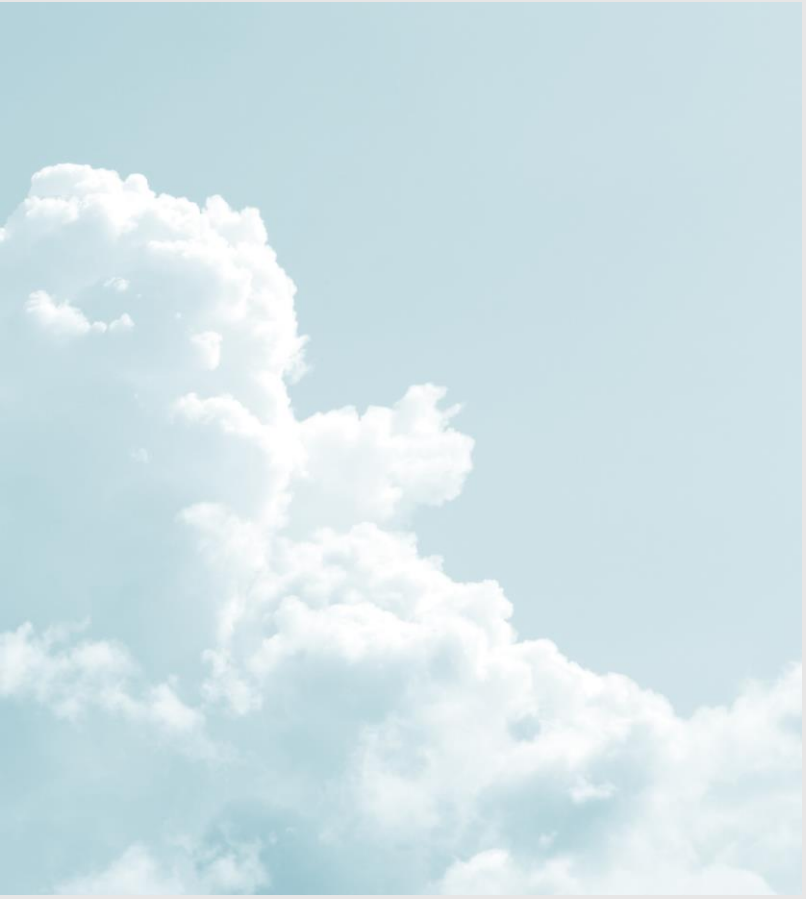
McMaster RDM Services has a **Data Storage Finder**, an interactive tool to help you find a vetted storage provider depending on risk, volume, and other needs.

This tool also allows you to compare feature sets of selected options.



https://rdm.mcmaster.ca/finder

Data Storage Platforms



Local Storage is any storage device directly present under your control: *laptops, desktop hard drives or SSDs, external hard drives, USB flash drives, or other storage media.*

Cloud Storage is a storage platform (typically run by a 3rd party) off-site that you access over the internet.

Somewhere between these are **Network Storage** devices; like university or department servers or Network Attached Storage (NAS) devices.

Local Storage - **Overview**

- Local storage is the most **common** storage option, used by many researchers.
- Every computer comes with some kind of local storage—a hard drive or SSD storage device—built in, making it very **convenient**.
- However, it relies on you to ensure that storage devices are properly **maintained, backed up, and secured**.
- Maintenance and backup should be considered as part of your **Data Management Plan**.

Local Storage – **Advantages**

- **Speed:** Faster access to data than cloud-stored data accessed over the internet or network drives.
- **Cost and Volume**: A 4 TB hard drive can be purchased for less than 100$.
- **Ease of Use**: Local storage devices are relatively easy to use and don’t need technical training.
- **Data Security**: Only the person who physically has the drives has access to the data. Drives can be encrypted easily.
- **Offline Access:** Data remains accessible even when the network or internet is not available. Drives are very portable and can be carried into remote locations

Local Storage – **Disadvantages**

- **Local storage devices are susceptible to data loss**
  - If a USB drive is unplugged from a computer while data is being transferred the drive can become corrupted.
  - If a hard drive is dropped it can damage the physical part of the storage media.
  - Small portable storage devices like USB keys and external hard drives are easy to lose or have stolen.

Organization of storage devices can become unwieldy as the number of devices increases.

r/GradSchool • 4 yr. ago
[deleted]

I lost 2 years' worth

The title says it all..

I don't know how to process this rig

I've been doing astronomy research
on the computer in my office, until

At the time I thought it was just a m
and I actually used this time to take

Yesterday they got back to me, and
campus, and somehow something w
scared" mode, and they told me the
do a clean restart.

They haven't done it yet, because th
do it, because I'd lose 2 years' worth

I already know people are gonna sa
but it just never crossed my mind (fu
scheduled for next Wednesday). I ha
large chunk of my work will definite

The projects that I was working on w

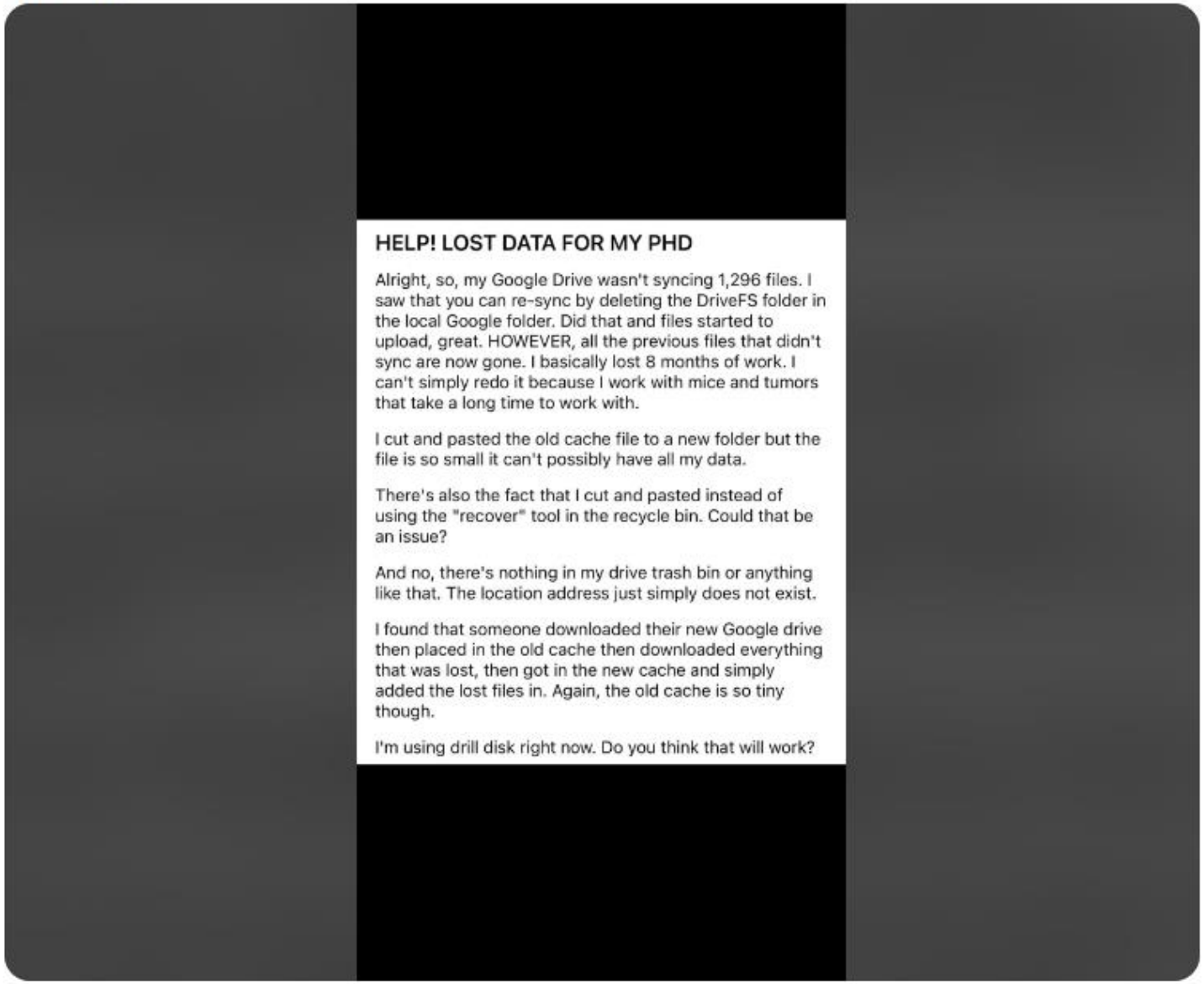
Read more

475 80

r/PhD • 1 yr. ago
JuriPlz

Lost a TON of data for my PhD and my PI is pissed. Does anyone have advice?

Need Advice



...

...

...y not checking if the

...nce I just defended my PhD today

...data, wrote the code, did all the

...l so they had to take my PC and wipe it
...l the files back.

...d send to me quickly.

...cup is done, and then I give it to them.

...). I just saw the tick mark, and in my
...handful of small files did get backed up,

...had a mini heart attack once I realized
...blished works, they were lost too.

...t wasn't published yet. Also, I actually
...rom memory, and thankfully I
...y work, but managed to replicate the

...it.

Local Storage – *Case Study: University of Manitoba*

- On March 28, 2009, a fire broke out in the Duff Roblin Building at the University of Manitoba.
- This fire impacted the Department of Psychology offices and laboratory space.
- For many researchers and graduate students, data was stored on desktop and laptop computers within the building.
- A Fire Recovery committee had to attempt recovery for the data located on local storage.
- Although most data was recovered, this is an excellent example of unexpected disasters – we can't rely on local storage alone!

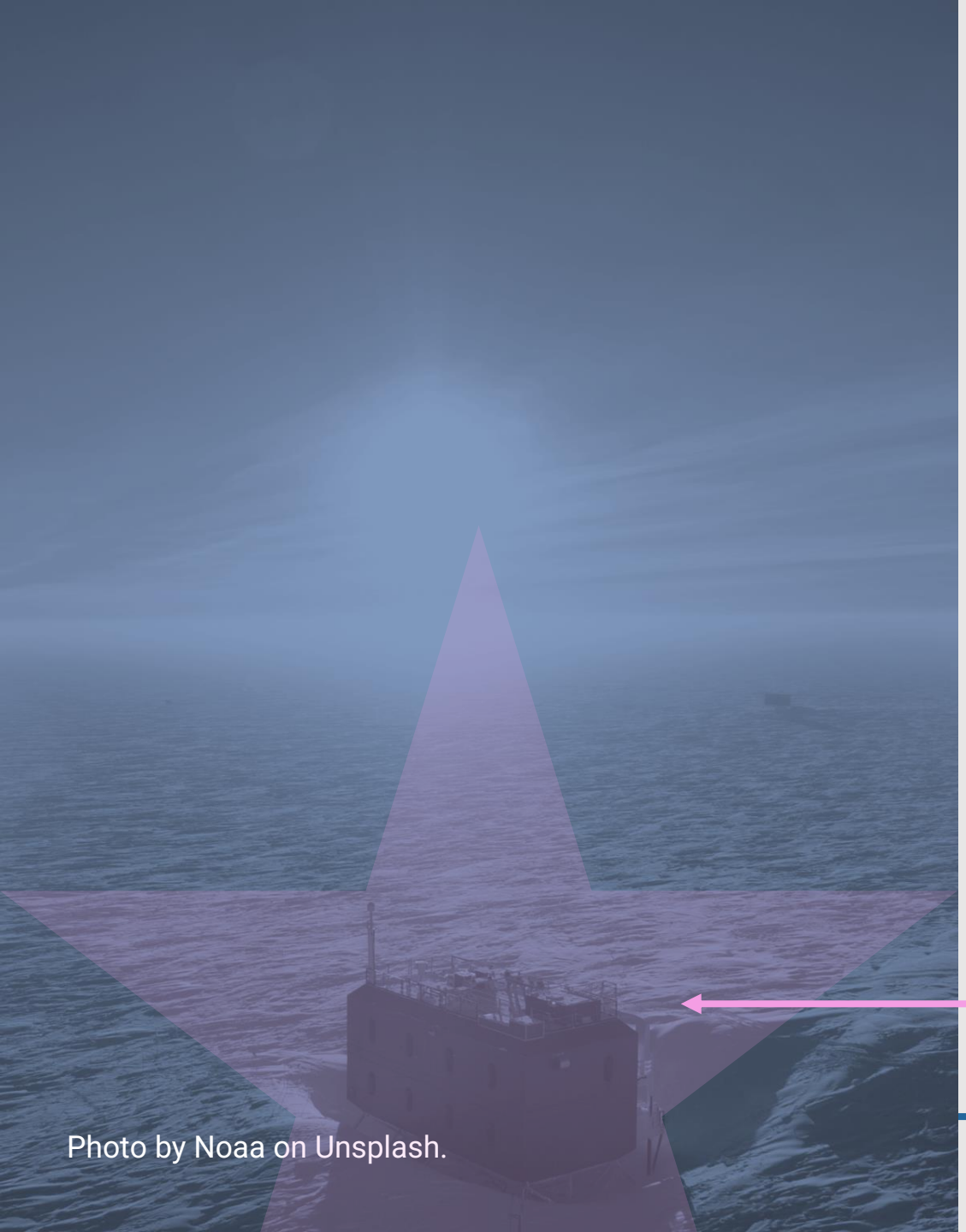


Photo by Noaa on Unsplash.

Local Storage - **Best Case**

Local storage devices are a good choice for:

- Data that is collected **remotely** in areas with no or limited internet connectivity.
- **Large amounts of data.**
- Data with **large amounts of local processing** required.
- **Sensitive data** that shouldn't leave the institution.

McMaster University | Library

Local Storage - **Hardware**

RAID - (Redundant Array of Independent Disks)

- Combines **multiple hard drives** to store data
- Data is distributed across the drives so that if one drive fails, data is not lost (RAID 1 to 6 standard).
- The downside of RAID devices is that more capacity is needed – 2 TB of storage might be needed for 1 TB of data.
- Data is still only stored in one location

Networked Storage - **Overview**

- **Department Servers:** Typically, will have regular backups, servers on campus. Storage volume and other details depend on the implementation.

- **Network-Attached Storage (NAS)**: NAS devices are small file servers that can be set up as a network drive accessible over the university or a separate local network.
  - NAS devices can often be set up as RAID devices.

Networked Storage – **Advantages**

- **Faster vs Cloud:** Faster than connecting via internet
- **Secure Access:** Access is limited to authenticated users on the network.
- **Easy to Use:** Appears as a drive on personal computers.
- **Central Storage:** All of research group's data can be stored in one place.
- **Automated Backups:** Department servers are backed up; NAS devices can be set up to automatically back up to some cloud storage providers.

Networked Storage – **Disadvantages**

- **Slower vs Local:** Slower than storage devices connected directly using ethernet, USB 3, or Thunderbolt.
- **Setup:** NAS devices may be difficult to set up and update – help from Departmental IT may be required.
- **Security:** NAS devices connected to the internet may be vulnerable to attack through bugs or if not kept updated.
- **Dependent on Power:** Power outages can affect devices and corrupt files if they are unprotected

Photo by National Cancer Institute on Unsplash.

Networked Storage - **Best Case**

Networked storage devices are a good choice for:

- Research groups that produce **large amounts** of data and want to store all their data in a **central location**.
- Data storage for **research instruments**.
- **Sensitive data** that shouldn't leave the institution.

McMaster University | Library

Networked Storage – **McMaster Institutional Providers**

Research and High-Performance Computing Support (RHPCS)

- High-performance computing and large data storage.
- 1 TB + with options for backup and servers.
- Cost recovery model – storage must be considered as part of grant applications and annual budgets.
- Data is stored on campus in the A.B. Bourns building
- Good for automatic backups of medium or high-risk data that should be kept on McMaster premises.
- RHPCS servers are useful for research groups needing large amounts of data storage for data that is actively being used.

Cloud Storage - Overview

Cloud storage providers have useful features like **automated backups**, **file versioning**, and easy **file sharing** between devices and users.

We can separate cloud storage into two categories:

- **Public:** available to anyone and a contract is made directly between you and the provider.
- **Institutional:** available to McMaster users and there is a contract between the cloud provider and McMaster.

Google Drive

Cloud Storage – **Advantages**

- **File Versioning and Recovery:** Changes to files are tracked and can be reverted. Deleted files can be recovered (for a certain amount of time).
- **File Synchronization:** Files are synchronized between the online platform and any linked devices including computers and mobile devices.
- **File Sharing:** Files can be easily shared with others, especially other users on the same platform.
- **Access from Anywhere (where there's internet)**

Cloud Storage – **Disadvantages**

- **Security:** Vulnerable to hack or through phishing etc. Storage locations are unclear. Data and metadata visible to cloud service employees.
- **Compromised Everywhere:** If your account or device is compromised, attackers can delete files; changes are synchronized across other devices.
- **Cost & Volume:** Ongoing subscription fees which increase with data volume.
- **Speed of Access:** Files that are not stored locally must be downloaded to access them.

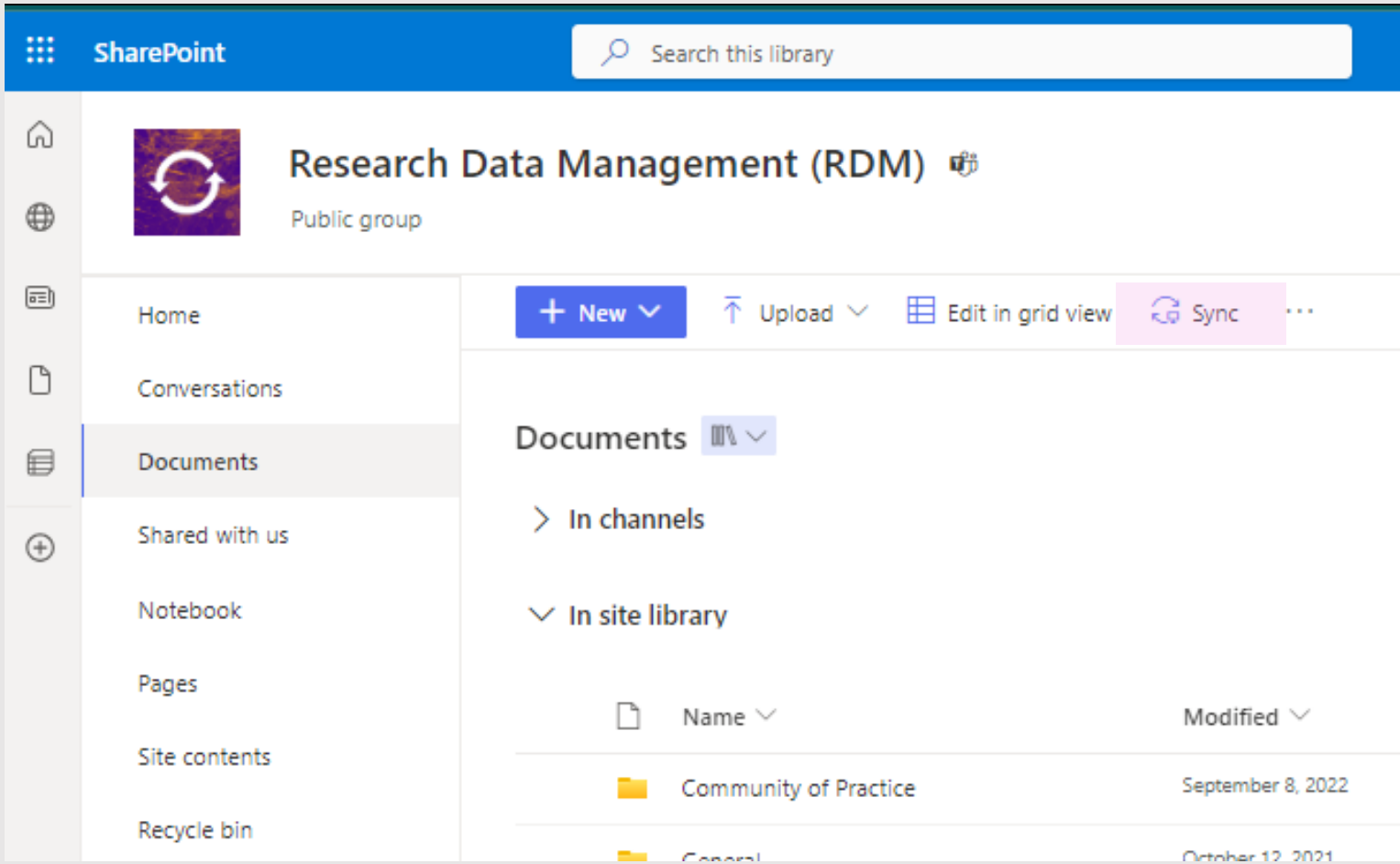
Cloud Storage - **McMaster Institutional Providers**

Microsoft Products – OneDrive, Teams, SharePoint

- Integrated with Microsoft Office products like Word & Excel
- Integrated with MacID – free for McMaster users
- Teams/SharePoint have excellent collaboration and user management features and are great for managing data for a research group
- Desktop and mobile applications available for MacOS, Windows, iOS, and Android
- Data is stored in Canadian servers.
- Data can be restored up to 30 days after deletion.
- OneDrive may be used for some medium/high-risk data if using encryption.
- **OneDrive**: 1 TB initial storage, up to 5 TB can be requested from UTS
- **Teams/SharePoint**: 25 TB of storage is available per Team/Site

Syncing Sharepoint

- Go to Team Files or Sharepoint
- Click “Sync”
- Files appear in “File Explorer”
- Pin to Taskbar
- OneDrive will also appear here



Cloud Storage - **McMaster Institutional Providers**

MacDrive

- Free for McMaster Users, integrated with MacID
- Applications available for MacOS, Windows, iOS, and Android
- 300 GB quota (can be increased by special request)
- Data is stored in servers on campus.
- Users can create encrypted folders.
- Upload-only public folders can be created (research participants/coordinators can share forms or data with researcher without being able to access other files)
- Students cannot make accounts directly, must be invited by faculty sponsor.
- May be used for medium/high-risk data if using encryption.

Cloud Storage – **Public Providers**

Dropbox, Google Drive, Box, Backblaze, etc.

- Pay per month, per user, and with increasing costs for storage.
- Data may be stored in the USA, or in Europe, or elsewhere
- No integration with University systems.
- Most have apps available for MacOS, Windows, iOS, and Android
- Feature sets vary – Dropbox/Google Drive have collaborative word processing.
- Integration with GitHub, OSF, and other platforms
- Privacy/Data security varies
- Potential service provider staff can look at your data. Risk of phishing attacks and hacks. Sometimes data can be lost due to syncing errors.

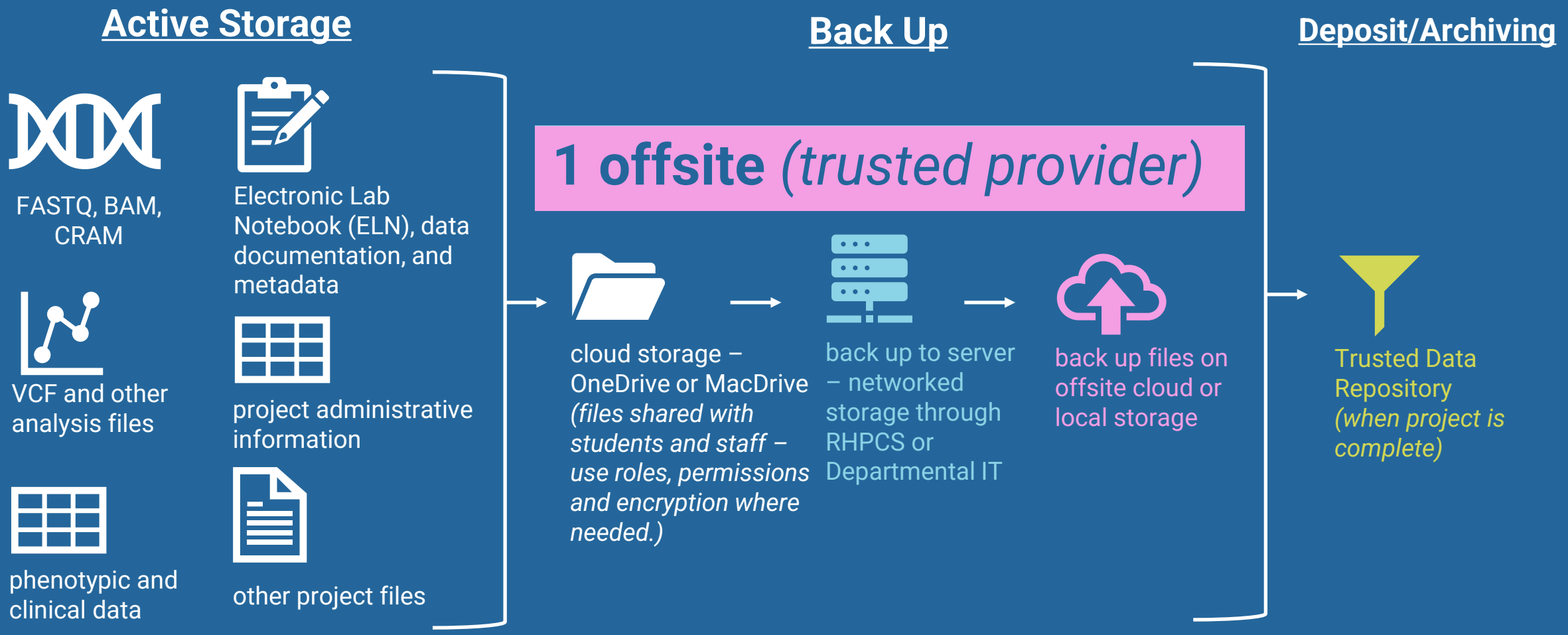
Photo by Imgix on Unsplash.

Cloud Storage – **Best Case**

Cloud storage is a good choice for:

- **Small-medium** amounts of data with **no special requirements**.
- **Collaborative** research with researchers split between different locations.
- Object storage can handle large amounts of data with large amounts of **scripted computing** processing required.

McMaster University | Library



File Storage Workflow – **Bioinformatics Data**

RDM Certificate Program

- Certificate you can add to your CV or ORCiD
- Attend 7 RDM workshops to receive a certificate!
- Go to this website to verify today's session: **https://u.mcmaster.ca/verification**
- Learn more about the Certificate Program: **https://scds.ca/certificate-program**

What is the Digital Research Alliance of Canada?

- https://alliancecan.ca/
- Started in 2019, it's a collection of universities in Canada that have access to certain *high performance computing systems*
- These systems include *large supercomputer clusters*, *research data management* solutions, and *research software*.
- There are two levels of access to the Alliance's HPCS service.

Rapid Access Services (RAS) – **Normal Usage**

- RAS access is granted to qualified staff from a member institution.
- RAS allow shell access and moderate storage (around 20TB scratch/1TB permanent, can vary by cluster).
- access to RAS Graphical Processing Units is allowed on an *opportunistic* basis.

Digital Research Alliance of Canada | Alliance de recherche numérique du Canada

Who qualifies?

- Generally, *faculty* or *librarians* from a member institution.
- ...but those people can *sponsor* other people.

Resource Allocation Competition (RAC) - **Allocations**

The Alliance's RAC is a *competitive* process for services – both storage and compute – beyond what's offered by RAS. It involves a fair amount of planning, paperwork. and a Canadian Common CV application, but secures you dedicated resources beyond what's accessible via RAS / normal usage.

This could include **much more** storage, or dedicated GPU/Compute, or other resources.

Digital Research Alliance of Canada | Alliance de recherche numérique du Canada

⚠ **Sensitive Data** – Storage Requirements

Research data can be separated into 3 levels of sensitivity:

- **Low risk** is research data that does not contain any sensitive or identifiable information.
- **Medium risk** is research data that may or does contain confidential, sensitive, or identifiable information *(Personally identifiable information, demographic data, etc.)*.
- **High risk** is research data that contains highly sensitive information *(Personal health information, personal financial information, sensitive ecological data, etc.)*.

Photo by Michael Dziedzic on Unsplash.

Sensitive Data

Medium risk or **High risk** data comes with special requirements for data storage.

- Data must not be stored on public storage services.
- Data must be **encrypted** when stored on a device connected to the internet.
- Data must be **encrypted** in transit from one device to another.
  - **High risk** data must not be sent via email or other public or unsecured methods
- Personal health data has even further requirements to comply with provincial health legislation (PHIPA in Ontario).
- Storing and working with sensitive data is a complex problem and we are available for consultations.

Encryption

Encryption is a process of transforming information so that it is only readable to a person with the correct authorization. Disk encryption can be implemented at a few different levels:

- **Individual Files:** Microsoft Office or other applications can be used to password protect and encrypt documents on a file-by-file basis.
- **Full Disk Encryption:** computers and mobile devices can have the whole disk encrypted.
- **Virtual Encrypted Disks:** A virtual disk can be created and then mounted similar to a USB key or other external drive.
- For more details see our webpage: https://rdm.mcmaster.ca/secure
- Join us March 20th for a session on encryption

Photo by Markus Spiske on Unsplash.

McMaster University | Library

Data Repositories: Taking care of data so you don't have to!

So far, we've only talked about data storage for active research projects. What is your plan for your data once your research project is finished?

Research data can be **published** or **archived** in an online **data repository**.

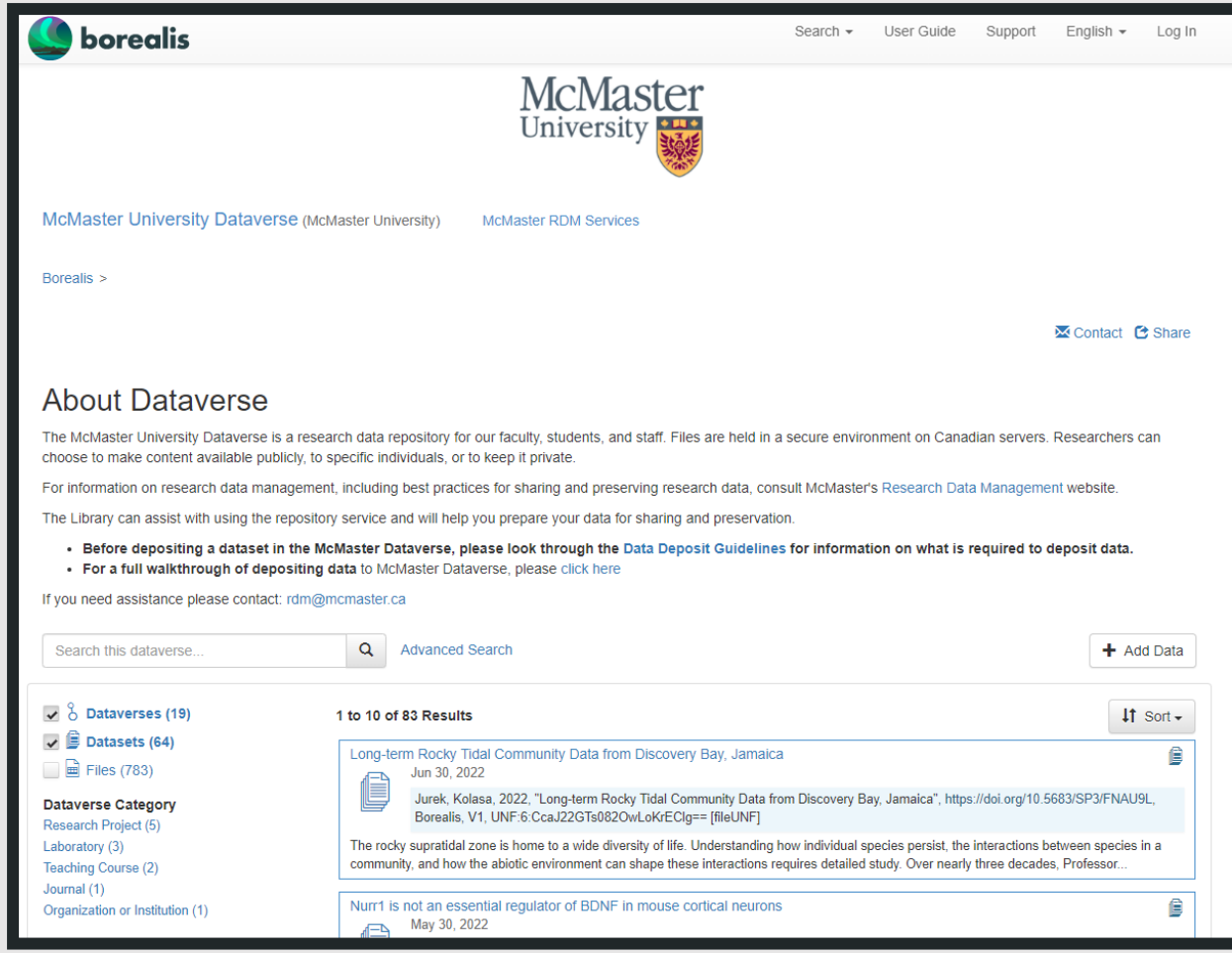
Research data repositories are the best way to keep data from being lost! *Storage devices can become corrupted or lost, you (or your research team) might move to a different institution.*

Photo by NASA on Unsplash.

McMaster Dataverse

https://borealisdata.ca/dataverse/mcmaster

- McMaster's Institutional Data Repository is a home for all research data originating from McMaster researchers.
- Provides basic data curation services
- Data is stewarded by professionals at McMaster
- Choose whether to share data openly or through an application process



RDM Services: **Community of Practice**

Teams channel hosting asynchronous discussion

Monthly meetings featuring presentations from researchers across the university on how they do data management.

February 29 – 11 AM: Data and Research Impact – Jack Young and National Collaborating Centre for Methods and Tools.

https://u.mcmaster.ca/rdm-community

March 20, 2024 | 10:30am-12pm
Virtual Workshop + Sandbox Session

How to Implement Encryption to Protect Your Research Data

u.mcmaster.ca/scds-events

RDM

SCDS

Library

McMaster University

April 17, 2024 | 10:30am-12pm
Virtual Workshop + Sandbox Session

Sensitive Data Management

u.mcmaster.ca/scds-events

RDM

SCDS

Library

McMaster University