



# Strategies for research data storage and backup

Isaac Pratt, PhD

June 3<sup>rd</sup>, 2021





- McMaster University sits on the traditional Territories of the Mississauga and Haudenosaunee Nations, and within the lands protected by

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



# SESSION RECORDING

- 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.

# 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/](https://scds.ca/events/code-of-conduct/)

# HELLO! A bit about me:

- I am a Research Data Management Specialist working in the University Library and Research and High Performance Computing (RHPCS).
- My background is in Biological Anthropology, Medical Imaging, and Human Anatomy.
- I have a PhD in Anatomy & Cell Biology from the University of Saskatchewan.
- Email me with any RDM question or to set up a consultation:  
[pratti@mcmaster.ca](mailto:pratti@mcmaster.ca)

# OUTLINE FOR TODAY

- Data storage principles
- Storage options for actively used data
  - Local storage
  - Networked storage
  - Cloud storage
- Data security precautions
- Sensitive data
- Storage options for publishing and archiving data

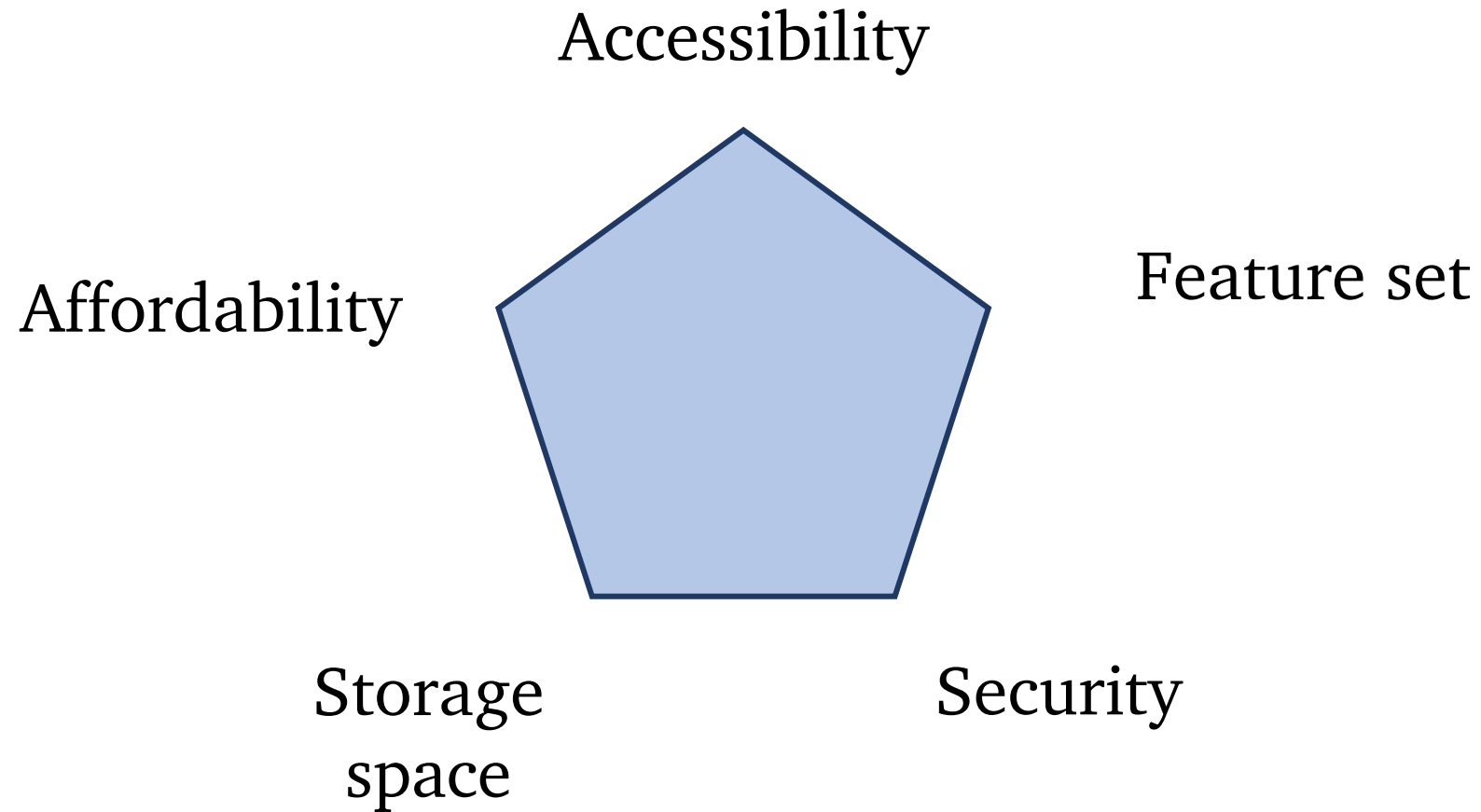
# DATA STORAGE PRINCIPLES

Every data set has a different set of specific needs and these must be matched with the feature sets of data storage platforms to find the best platform for the data.

These needs can include:

- Large storage requirements – with regular increases
- Sensitive data storage
- Computation/analysis needs

# DATA STORAGE PRINCIPLES





# DATA STORAGE PRINCIPLES

In general:

- More secure = Less accessible

How secure does your data need to be?

What other features are you looking for?

integrated,  
location, control,  
collaborative,  
reliability,  
shareability,  
Speed,

# BACKUP STRATEGIES

Whatever storage solution you choose, it's important to have a backup strategy.

3

Copies of your data

2

Copies are on-hand (easily accessible)

- a “**production**” (working) copy
- a “**production backup**” copy

1


Copy is in another location (“off-site”), with a *trusted* service provide



# RESEARCH DATA STORAGE FINDER TOOL

Step 1: Answer these questions to narrow down storage provider options.

**CLEAR ANSWERS**

1. What risk level is your data?  


Low  
 Medium  
 High

2. What type of data storage are

Step 2: Select data storage providers you would like to compare

**SELECT ALL** **CLEAR SELECTIONS**

<b>Compute Canada</b> Advanced research computing systems, storage and software	<b>Compute Canada NextCloud</b> Advanced research computing File hosting services	<b>Dataverse</b> Store, share, publish and discover research data
<b>FRDR</b> Find and Share Canadian Research Data	<b>Github</b> Distributed version control system for software code	<b>MacDrive</b> File Synchronization and Sharing solution
<b>MacDrop</b>	<b>McMaster-based</b>	<b>OSF</b>

<http://u.mcmaster.ca/storagefinder>

# DATA STORAGE PLATFORMS

## Definitions:

- **Local storage** is any storage device directly present under your control.
  - Includes things like laptop or desktop hard drives or SSDs, external hard drives, USB flash drives, DVDs or other storage media, etc
- **Cloud storage** is a storage platform (typically run by a 3<sup>rd</sup> party) off-site that you access over the internet
- Somewhere between these are **Network storage** devices like university or department servers or NAS devices.



# LOCAL STORAGE

- Local storage is the most common storage option
- Local storage is very convenient, but relies on you to ensure that storage devices are properly maintained, backed up, and secured.



# LOCAL STORAGE ADVANTAGES

## Speed

- Local storage provides much faster access to data than cloud stored data accessed over the internet.

## Cost & Volume

- A 2 TB hard drive can be purchased for less than 80\$ and costs scale well.

## Ease of use

## Data security

## Portable

## Offline access

- Local storage devices remain accessible even when the network or internet is not available.



# 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 while active 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 be stolen.

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

# University of Manitoba Psychology



National M



Winnipeg Free Press



STRAFFORD AGENT IMAGES



April 19th  
s stolen  
tore my  
e to say  
pay you  
E-YEAR  
a folder  
THESIS  
which is  
and use  
price is  
address

and I would appreciate it so so



# LOCAL STORAGE

**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
- Small volumes of data with large amounts of processing required
- Sensitive data that shouldn't leave the institution



# LOCAL STORAGE

## 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.
- The downside of RAID devices is that double the capacity is needed – 2 TB of storage is needed for 1 TB of data.
- Data is still only stored in one location



# NETWORKED STORAGE

## Department servers

- Typically will have regular backups, servers on Campus. Storage volume and other details depend on the implementation

## NAS “Network-Attached Storage”

- NAS devices are small file servers that can be set up as a network drive accessible over the university or a separate local network.
- Can be set up as RAID devices.



# NETWORKED STORAGE ADVANTAGES

## Speed

- Faster than connecting via internet, slower than directly connected

## 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 a 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

## Speed

- Faster than connecting via internet, slower than directly connected

**NAS devices may be difficult to set up**

**Power outages can affect files**

# NETWORKED STORAGE

**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

# CLOUD STORAGE

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



Google Drive

- We can separate cloud storage into two categories:
  - **Public** cloud providers are available to anyone and a contract is made directly between you and the provider.
  - **Institutional** cloud providers are made available to McMaster users and there is a contract between the cloud provider and McMaster.





# 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

- Your cloud account may be compromised through a hack or through phishing or another technique.
- Data storage locations for many cloud services are unclear.
- Data and metadata may be visible to employees of the cloud service.

## Comprised somewhere is compromised everywhere

- If your account or one of your devices is compromised attackers can delete files and those changes can be synchronized with your other devices.

## Cost & volume

- Most public cloud services charge ongoing subscription fees which increase with data volume.

## Speed of access

- Files that are not stored locally must be downloaded to access them.

# MCMASTER INSTITUTIONAL CLOUD STORAGE PROVIDERS

## Microsoft OneDrive

- Integrated with Microsoft Office products like Word & Excel
- Integrated with MacID
- Applications available for MacOS, Windows, iOS, and Android
- 1 TB initial storage, up to 5 TB can be requested from UTS
- OneDrive data is stored in Canadian servers
- Data can be restored up to 30 days after deletion.



# MCMASTER INSTITUTIONAL CLOUD STORAGE PROVIDERS

## MacDrive MacDrive

- Integrated with MacID
- Applications available for MacOS, Windows, iOS, and Android
- 300 GB quota
- Data is stored in servers on campus.
- Users can create encrypted folders.
- Upload only public folders can be created.
- Students cannot make accounts directly, must be invited by faculty sponsor.

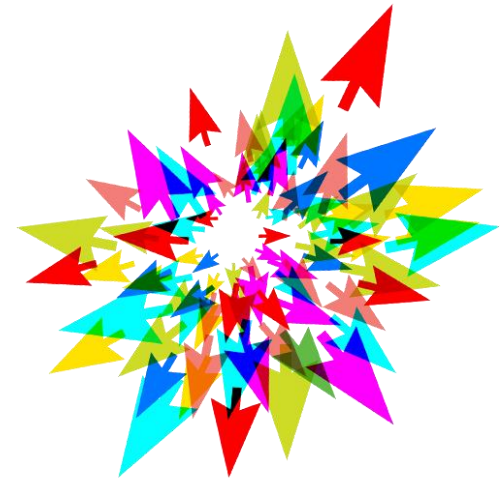


# MCMASTER INSTITUTIONAL CLOUD STORAGE PROVIDERS

## Compute Canada

- High performance computing and large data storage
- Initial allocation of 1 TB can be increased up to 10 TB and beyond
- No cost for users, PI must make initial account.
- Data is stored in Canadian servers.
- Massive amounts of computation power can be used to analyze data.

**compute** | **calcul**  
canada | canada



# PUBLIC CLOUD STORAGE 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
- Privacy/Data security varies

# OBJECT STORAGE PROVIDERS



**Amazon S3, Microsoft Azure, Google Cloud Storage, IBM Cloud Storage, Backblaze, etc**

- Object storage manages data as objects with metadata rather than in hierarchical folder based systems.
- Object storage systems are best for massive amounts of unstructured data.
- Most providers have an unlimited storage capacity but you will typically pay not only for storage but also each time you access/retrieve the data files.
- Object storage systems typically come with computing ability
- Great for large unstructured databases and data sets

# CLOUD STORAGE

**Cloud storage is a good choice for:**

- Small-medium amounts of data with no special requirements
- Collaborative research with researchers split between different locations.
- Large amounts of data with large amounts of scripted computing processing required



# CLOUD STORAGE SECURITY

- **Password strength** increases with:
  - Length (use 8+)
  - Use of both lowercase and uppercase letters
  - Use of numerals
  - Use of symbols
- Try using a memorable phrase to make passwords easier to remember:
  - Instead of aE8\*ngh789@ use M@yfa1rCrescent
- Or use a **passphrase** with words and spaces
  - calibrate hunting stole oval
- Use a unique password for each important service
- Use a **password manager** such as LastPass, 1Password, or KeePass
- Check password strength: <https://www.uic.edu/apps/strong-password/>

# CLOUD STORAGE SECURITY

## Multi Factor (2 Factor) Authentication

- A website will ask for a second authentication code from a user trying to login.
- The 1st Factor is your password
- The 2nd Factor is a security code sent via text message or email or obtained from a linked authenticator app
- MFA increases the security of your account because an attacker will need to access your phone as well as having your password.
- MFA can be enabled for your McMaster Microsoft account here <https://office365.mcmaster.ca/mfa/>

# SENSITIVE DATA

## McMaster defines 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

# 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. Encryption can be implemented at a few different levels:

- **Full Disk Encryption:** computers and mobile devices can have the whole disk encrypted.
- **Virtual disk encryption:** 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://library.mcmaster.ca/services/rdm#tab-secure-your-data>

# PUBLISHED DATA

So far we've only talked about data storage for active research projects. What do I do with my data when my research project is finished?

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



# PUBLISHED DATA



The new Tri-Agency policy will require research data supported by Tri-Agency grants to be **deposited** in a data repository. (Not required to be openly shared)

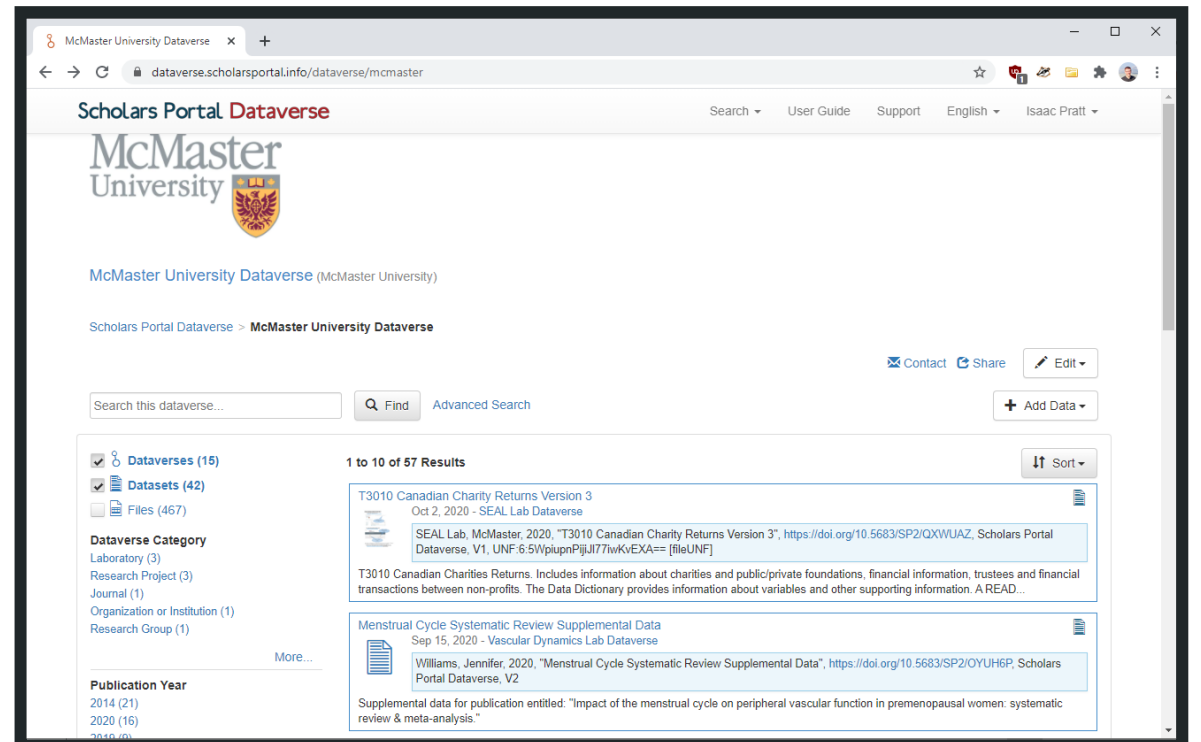
Depositing data is great way to archive finished research data and make it available to other researchers:

- Increases research reproducibility and validation.
- Increases the impact and visibility of research through re-use and linkage to other studies.

# MCMMASTER DATAVERSE

McMaster Dataverse is our Institutional Data Repository  
<https://dataverse.scholarsportal.info/dataverse/mcmaster>

- Built for datasets
- Contains tools for tabular data exploration and analysis
- Allows researchers to control how they license and share their datasets





# PUBLISHED DATA

Our next webinar will be on publishing and sharing data:

[Essentials of open data sharing](#) on Thursday, July 8, 2021 at 2 pm

Research Data Management Summer Series

## Essentials of open data sharing

Synchronous  
Workshop

July 8, 2021 | 2:00 pm

McMaster University | Research & High Performance Computing | McMaster University | Library

McMaster  
University

Research &  
High Performance  
Computing

McMaster  
University

Library

# THANK YOU!

**For more information:**

Visit: [library.mcmaster.ca/services/rdm](https://library.mcmaster.ca/services/rdm)

Check out the whole webinar series at:  
<https://scds.github.io/intro-rdm/>

Contact us at: [rdm@mcmaster.ca](mailto:rdm@mcmaster.ca)

**RDM**  
**@McMaster**