Storage, backup, and security

Objectives

- Understand risks associated with data storage
- Understand data back up
- Design a storage strategy

Introduction

In this module, we'll focus on storage of digital data. We'll cover storage locations, and discuss data backup and security. The goal is to think "safety first," and reduce risks.

What risks?

We'll focus on reducing the likelihood of:

- Data loss
- File corruption
- Unauthorized data access

Data loss may occur if a file is accidentally deleted, a computer is stolen, or a hard drive fails and data are not backed up.

File corruption may occur when digital files are moved or copied, and when they are stored on one medium for a long period of time.

Unauthorized access may occur if appropriate security precautions are not followed for sensitive data.

Discussion: Choosing storage and back up locations

Group discussion, taking live notes for each storage location

The location of your data will likely depend on the degree of freedom you have in the management of your data. Your data may need to be stored in a particular location due to legal and ethical considerations, or due to your lab or research group's practices.

In this section we will survey 4 types of storage locations, discussing features that help to assess the benefits and risks of using particular locations for storing and backing up your data.

We'll talk about:

- Accessibility
- Space
- Security
- Integrity

Personal computers

- Recommended for working with data that aren't sensitive or restricted.
- Depend heavily on personal practices.

Accessibility

- Convenient for storing data while in use.

Space

- Limited to space available on the computer hard drive.

Security

- Can be lost or stolen.
- Can have multiple valid users with access.
- Hard drive may or may not be encrypted.

Integrity

- Files are stored on the local hard drive, which will eventually fail.
- Depends on personal practices for:
    - Regular backup to another location.
    - Software updates.

External hard drives or USB sticks

- Usually used as a location for backup or for data sharing.
- Depend heavily on personal practices.

Accessibility

- Low cost
- Portable, can be kept locally and attached to personal or work computers

Available space

- Same scale as personal computers, limited to space available on the hard drive or USB stick.

Security

- May easily be lost or stolen.
- Hard drive or USB stick may or may not be encrypted.

Integrity

- USB sticks are flimsy.
- External hard drives will eventually fail.

University networked drives

- Recommended for storing master copies of data.
- Managed by professional IT staff.

Accessibility

- Made accessible via log in on personal or work computer.

Available space

- Depends on agreement with IT, may have a cost.
- Capable of storing large datasets.

Security

- Access controlled by IT staff.
- Measures may be set up for storing sensitive and controlled data.

Integrity

- Storage is distributed, with redundancy for failure of a single hard drive.
- Regular update and back-up strategies are in place.

Third-party services

TAMU-sponsored services:

- TAMU Google Drive
- Syncplicity

Other common services:

- Amazon S3 ("cloud" computing storage facilities)
- Dropbox
- SpiderOak

Accessibility

- Accessible via log in, usually via web browser or a computer application.
- Store files remotely and synchronise files across local computers.
- Most are platform-agnostic.

Available space

- Usually provide the first few gigabytes free and users pay for more space.
- Capable of storing large datasets.

Security

- Depends on the service.
- Some services may have measures for storing sensitive and controlled data.
- Services operated on servers physically located outside US may be a problem for some data.

Integrity

- Usually provide versioning, but this may be limited.
- Storage is distributed, redundancy for failure of a single hard drive.
- Regular update and back-up strategies are in place.
- Service provider may go out of business.

TAMU-sponsored online services

Google Drive

Google Drive is a cloud storage service that allows you to store an unlimited amount of files that you can access anywhere you have an internet connection. You can upload documents, presentations, pictures, videos and anything else you may need.

Google Team Drive is a new addition similar to Google Drive, but is especially designed for teams. Files in Google Team Drive belong to the team instead of an individual user, so shared files remain even if a member leaves.

Users of the Service must follow both Google's Acceptable Use Policy and Texas A&M University IT Security Controls.

The Service is not appropriate for:

- Electronic Protected Health Information (EPHI) subject to the Health Insurance Portability and Accountability Act (HIPAA).
- Data controlled for export under Export Control Laws (EAR, ITAR).
- Certain types of Personally Identifiable Information (PII), including Social Security Numbers, credit card numbers, and bank or financial account numbers.
- High Risk Activities such as those involving business records in which loss or inappropriate disclosure would result in large consequences in terms of economic loss, loss of trust, or legal liability.

Data Sharing: Users of the Service can allow others to access data for the purpose of sharing and collaboration. Do not share data with anyone who does not have the appropriate authority to view it unless the data is of a public nature.

Google Apps at Texas A&M. "Terms Of Use and Privacy" **website**

Syncplicity

Syncplicity offers a secure cloud environment for users in The Texas A&M University System to store documents. It offers users the ability to sync any folder or desktop, include and exclude subfolders, and use native clients for Windows, Mac, iOS, and Android.

TAMU IT **website**

Example: Professor Richard Rodger

Watch 2-minute video

Backing up data with Professor Richard Rodger, University of Edinburgh **youtube**

Exercise: Thinking about storage

- Where do you think you will you store your data?
- How many copies of your data will you keep?

Storing data

Good practices for storing data to reduce the likelihood of data loss and file corruption.

1. Keep at least 3 copies of data in at least 2 geographically distributed locations.
2. Original
3. External backup, kept locally
4. External backup, kept at a remote location

For storage that you control, redundancy is an integral part of data management. Geographically distributed copies reduce the risk posed by a calamity at one location. You can think of it as insurance in case of theft, power outage, flood, fire, etc.

1. Check data integrity after copying data files, and regularly check backup files.
2. Restore your data files from backups to check that they can be read
3. Generate checksums

Digital files can be corrupted when they are copied, moved, or stored for a long period.

1. Copy data files to new media 2–5 years after first created.

All storage media (e.g. hard drives) break down over time.

Create a backup system

- How much space will you need to store your original data?
- Where you store your original data?
- How often should backups be made? A daily, weekly, or monthly schedule?
- Do you want a full backup or an incremental backup with changes on this schedule?
- How long should each backup be stored before being over-written by a new one?
- How much additional space will be required to maintain these backups?
- How will you keep track of different versions of data when backing up to multiple devices?
- Where will you backup your data locally? And remotely?

Use software to help

Once you have an idea of what you want. Software that automatically backs up files can simplify the process considerably.

These software tools can backup to the locations you choose, at the schedule you choose.

Show one

- Time Machine: Mac-only
- Arq: Mac, Windows (**website**)
- Crashplan: Windows, Mac, Linux (**website**)

Test your backup system

To be sure that your backup system is working, periodically retrieve your data files and confirm that you can read them.

You should do this when you initially set up the system and on a regular schedule thereafter.

Security tips

As we discussed in the "Legal and ethical considerations" module, if you're working with sensitive data, you're responsible for maintaining confidentiality of identifiable data.

This means data on your computer may need to be kept safe. Or you may be given access to a secure system that stores the data. Make sure to work with your advisor and your research team to understand and meet data security requirements.

General good practices to reduce the likelihood of unauthorized data access.

- Be aware of policies for moving or sharing the data.
    - Speak with your advisor.
- Avoid logging in to secure spaces using untrusted computers or networks.
    - Such as café wifi.

- Make sure physical storage media are in a locked room or safe
    - Lock up your computer or hard drives when they are not in use.
    - Servers managed by the university or third parties should also be located in secure rooms.
- Use strong passwords and be careful with your login credentials.
- Store and transmit sensitive data using encryption.
- Ensure complete destruction when deleting sensitive data.

Passwords

Tips:

- Create unique passwords for your accounts
- Use 18 or more characters where possible, and max it out if you can
- Use nonsense "passphrases" with fake words, numbers, and symbols

A password manager can make life easier:

- 1password
- LastPass

Encryption

Encryption is the process of converting data into an unreadable code. You must have access to a password or a secret encryption key to be able to read an encrypted file.

Encrypting your data will help ensure your data remain safe from disclosure in the event that a laptop, desktop, USB stick, or external hard drive are lost or stolen.

However, data encryption is not a substitute for other information protection controls. It still relies on the creation of a strong password. If the encryption key is lost, the disk image gets corrupted, or the hard disk fails, any encrypted data will be lost.

Use mainstream encryption tools.

Data files on computer You can encrypt individual files or folders on your computer.

Data files in transit Pretty Good Privacy (PGP) is a data encryption and decryption computer program that provides cryptographic privacy and authentication for data communication. PGP is often used for signing, encrypting, and decrypting texts, e-mails, files, directories, and whole disk partitions and to increase the security of e-mail communications.

Entire computer hard drive or USB stick

software:

- BitLocker: Windows full disk encryption
- FileVault: Mac

Fun fact: Login passwords on your computer are easy to bypass

Your computer password doesn't necessarily keep your data safe. Unless your files are encrypted on your computer, someone who has physical access to your computer can get access to your data with enough effort.

How? Basically, your computer's hard drive can be used like a big USB stick. File permissions for a user are set by your operting system. If someone boots your hard drive with their operating system, those permissions are irrelevant. This can involve different strategies for Macs and PCs, but in both cases it's easy to do with a bit of effort.

As a precaution, encypt your hard drive.

Find out More:

- PC example: hard drive taken out of computer and mounted on another host gives them access.
- Mac example: Connecting your Mac to another Mac and booting your Mac in Target Disk Mode, your Mac's hard drive will appear as a mounted drive on the other Mac. Unless Mac1 is encrypted with FileVault, all files will be accessible **website**

Sentitive Data Deletion

File deletion is not enough to ensure that sensitive electronic data are completely removed from a computer system. File deletion removes only the pointers to the disk sectors in which the data reside.

Deleted files can be recovered using commonly available software tools.

To ensure the complete destruction of sensitive data, the main options are:

- Data erasure, also called "data clearing" or "data wiping." This involves using software to remove all data by overwritting it. It leaves the hard drive or other storage medium operable.
- Degaussing. This disturbs the magnetic alignment of magnetic storage media and in many cases makes newer media, such as hard drives, unusable.
- Physical destruction. Through disintegration, shredding, pulverizing or incineration.

Conclusion

If you store your data on the university networked drives your data will be stored in a single place and backed up regularly. The risk of loss, theft or unauthorised use will be minimised as the data will be stored securely. Keep at least one copy of your master data files on the networked services.

Keeping backups is an essential data management task. There is a real risk of losing data through hard drive failure or accidental deletion. Recommended that you keep at least 3 copies of your data on at least 2 different media, keeping storage devices in separate locations with at least 1 off-site, and check that they work regularly. You should also have a policy for maintaining regular backups.

Bonus

- Before the next session, ask your advisor where he/she stores his/her data and how it's backed up
- If they use network drives, who manages them and do they know the security and backup policies?

References

- DMPTool. "Data Management General Guidance" **Website**
- Hicock, Robyn. 2016. "Microsoft Password Guidance" **PDF**
- MANTRA. "Storage and security" **Module**
- New England Collaborative Data Management Curriculum. "Module 4: Data Storage, Backup, and Security" **Website**
- Google Apps at Texas A&M. "Terms Of Use and Privacy" **Website**

Materials

Handout

Physical handout and available online for download

Questions to answer to build a storage and backup strategy:

- How much space will you need to store your original data?
    - How large are your data files?
    - Is your data collection iterative?
    - How much data will you accumluate every day, week, or month?
    - How much data do you anticipate collecting and generating by the end of your project?

- How do you want to back up your data?
    - Will all the data be backed up each time (full backup), or only amended data and changes (incremental backup)?
- How often will backups be made?
    - How often do you make changes to the data?
    - Should back ups be scheduled daily, weekly, monthly?
- How long will each backup be stored before being over-written?
    - For example: under the Grandfather-Father-Son rotation scheme, files may be available for two to three months before the space is over-written.
- How much additional space will be required to maintain the backups?
- How will you keep track of different versions of data when backing up to multiple devices?
    - Will you use a Version Control Systems (VCS)?
- Where will you backup your data locally? And remotely?
    - What hardware and services are available that meet your accessibility, security, and space needs?

Software list: