Data backup, security, storage & preservation

GEO 802 Fall 2020, Data Information Literacy

Anna C. Véron, Dr. sc. nat.

Lesson 7: Data backup, security, storage & preservation

→ **Why?**

→ **Where to store data**

→ **Data backup**

→ **Data security**

→ **Data preservation**

Why caring about data storage and preservation?

Swiss universities are guided by the research integrity guidelines of the Swiss Academies of Arts and Sciences.

Akademien der Wissenschaften Schweiz
Académies suisses des sciences
Accademie svizzere delle scienze
Academias svizras da las scienzas
Swiss Academies of Arts and Sciences

Die Projektleitung ist dafür verantwortlich, dass Daten und Materialien nach Abschluss des Projektes während einer für das Fachgebiet adäquaten Dauer aufbewahrt bleiben. Sie hat für ihre Haltbarkeit und Sicherung zu sorgen.

Research sponsors such as the SNF or Horizon2020 are increasingly imposing strict requirements on the storage of research data.

FNSNF
SCHWEIZERISCHER NATIONALFONDS
ZUR FÖRDERUNG DER WISSENSCHAFTLICHEN FORSCHUNG

Der SNF hat die Archivierung von Forschungsdaten in seiner Grundsatzerklärung zu Open Research Data festgehalten. Einen genauen Zeitrahmen definiert der SNF jedoch nicht, da dies je nach Disziplin oder Forschungsgegenstand variieren kann. Der SNF empfiehlt Forschungsdaten in der Regel für eine Dauer von 10 Jahren zu archivieren.

Why caring about data storage and preservation?

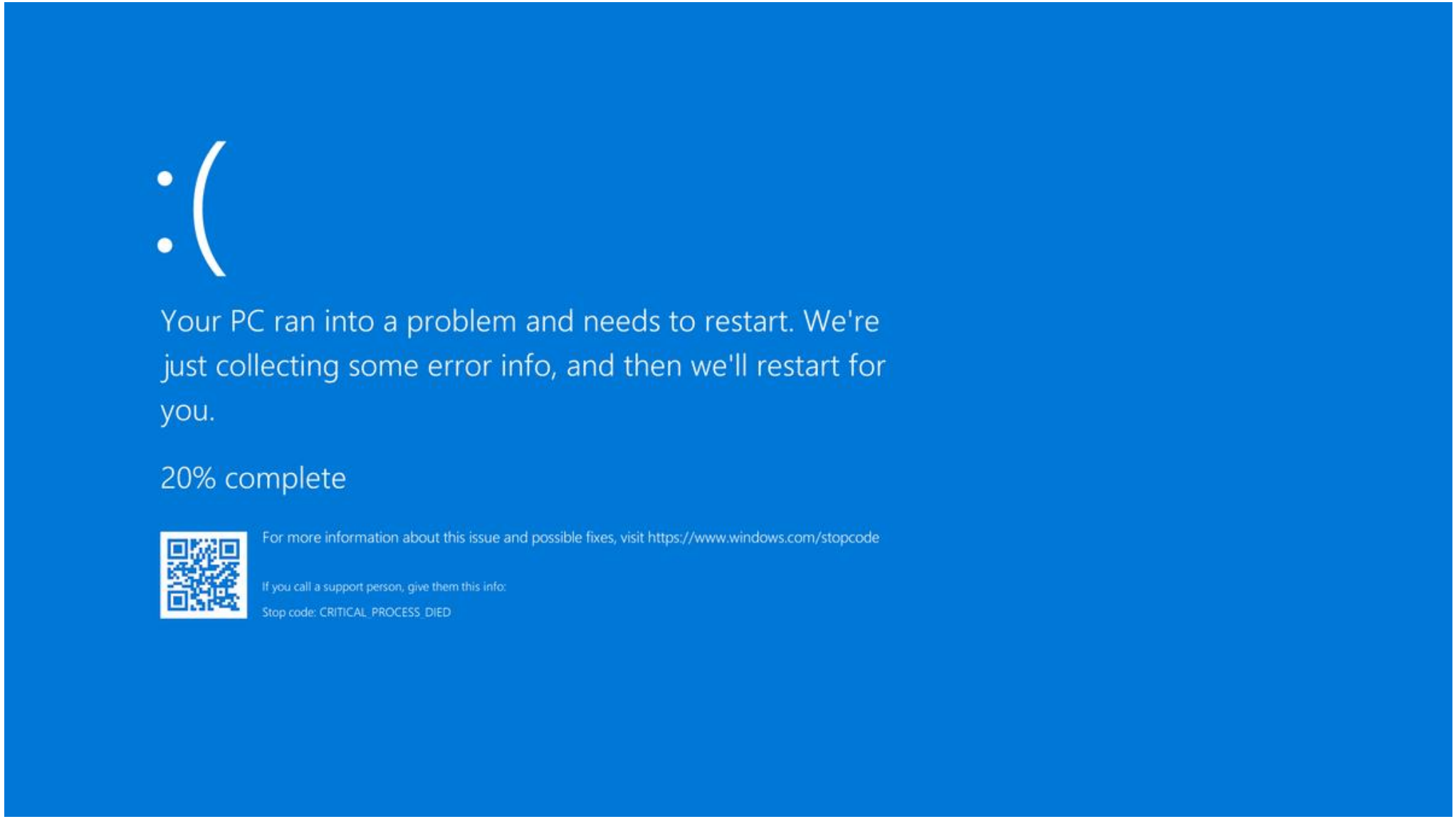
- Media formats age rapidly.
- Storage structures are not documented because they are clear to the person responsible.
- Contact persons are mobile and may only stay at the university for a few years.
- Storage locations are subsequently moved and links broken.
- The variety of subject-specific file and metadata formats makes long-term technical usability more difficult.

Image: Janet McKnight via Flickr, 15422638442 CC BY, icons by icon8.com

“Your data are the life blood of your research. If you lose your data recovery could be slow, costly or even worse…

it could be impossible.”

Blue screen of death…



University of Southampton, School of Electronics and Computer Science, Southampton, UK, 2005

Data loss will happen to you

Hard drive failures

Theft or loss of equipment

Dropping your laptop

Research trends (follow the money consequences)

File formats not readable anymore

People move to a new lab

Overwriting data

Media degradation (CDR's, memory sticks, hard drives, etc.)

Obsolescence / upgrades

Non-understandable data (bad metadata)

Lesson 7: Data backup, security, storage & preservation

- ✓ Why?
- → Where to store data
- → Data backup
- → Data security
- → Data preservation

Where do you store your data?

- PCs & Laptops
- External storage devices
- Network drives
- Cloud servers

→ Go to www.menti.com and use the code **11 12 83**

Storage: PC / Laptop

Pros:

- Convenient
- Accessible

Cons:

- Drive failures are common
- Susceptible to theft and damage
- Not replicated

Bottom Line: Do not use to store master copies of data. Only for «work in progress» files.

Storage: External storage devices

Pros:

- Convenient
- Cheap & portable

Cons:

- Longevity not guaranteed
- Easily damaged, misplaced or lost
  Security risk!
- Might not be big enough to hold all data

Bottom Line: Do not use to store master copies of data. Not recommended for long-term storage.

Storage: Network storage devices

Pros:

- Replicated storage: Less vulnerable to loss due to hardware failure
- Secure storage minimizes risk of loss, theft, unauthorized use

Cons:

- Costs
- Need network connection to access files
  (can be slow, especially when working on files with software)

Bottom Line: Highly recommended for master copies of data. Suitable for long-term storage (5 years or more).

Storage: in the Cloud

Pros:

- Backed up regularly and automatically
- Replicated storage: Less vulnerable to loss due to hardware failure
- Most provide versioning and encryption
- Secure storage minimizes risk of loss, theft, unauthorized use

Cons:

- Stored data *may* not be entirely private
- Service provider may go out of business --- Longevity?
- Possible restrictions by funding agency

Bottom Line: Recommended for master copies of data. Long-term storage?

Exercise 7.1: SWITCH drive

Try SWITCH drive for data entry and collaboration

1. Work in groups of three.
2. Create an account in SWITCH.
   https://preview.tinyurl.com/switchdrive
3. Create a file and share it with your partner.
4. Try to both work on it simultaneously.

Lesson 7: Data backup, security, storage & preservation

- ✓ Why?
- ✓ Where to store data
- → Data backup
- → Data security
- → Data preservation

Exercise 7.2: Backup vs. Archive

- What is Backup and what is archiving?
- Quickly note down your answer:
  Backup = x,y,z,a
  Archive = b,c,d,e

1 Data may change

2 Finalized data; static record

3 Kept long-term (5+ years)

4 Often stored in official archive

5 Preservation formats

6 Not permanent

7 Usually stored locally (individual, department, college)

8 “Working” formats

Exercise: Backup vs. Archive

Backups vs. Archiving

- **Backups**
  - Used to take **periodic snapshots of data** in case the current version is destroyed or lost
  - Backups are **copies of files** stored for short or near-long-term
  - Often performed on a somewhat **frequent** schedule
- **Archiving**
  - Used to **preserve data** for historical reference or potentially during disasters
  - Archives are usually the **final version**, stored for long-term, and generally not copied over
  - Often performed at the **end of a project** or during major milestones

- It is a good idea to have multiple copies of your backups and archives, in case one copy fails.

Backups: Considerations

- **Are there existing policies that might affect how and when you do data backups?**
  - May be separate project, office, department, funding source, or organizational policies
  - Are backups already part of a larger data management or contingency plan for your group?
- **Who is responsible for performing backups?**
  - Users?
  - System administrators?
  - Both?
- **Do the policies fit your needs?**

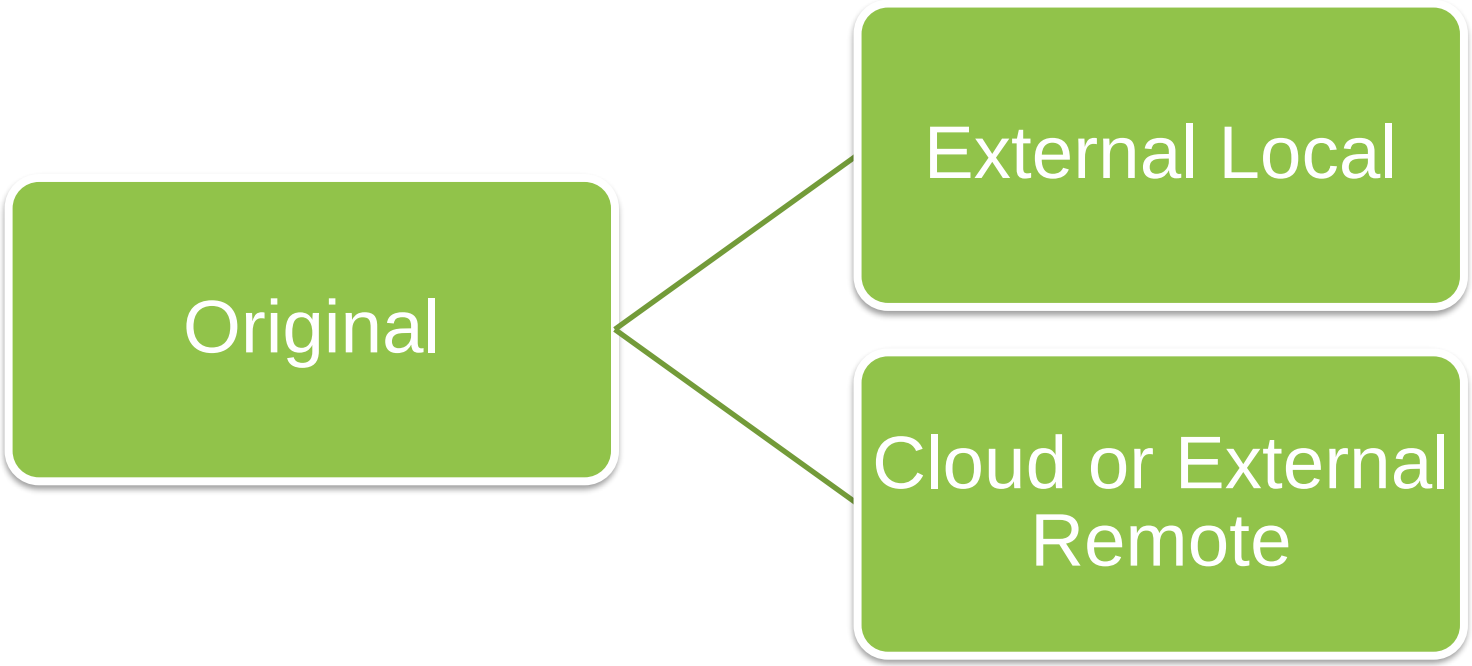
Backups: Considerations

- **How often should you do backups?**
  - Continually? Daily? Weekly? Monthly?
  - Cost vs. benefit
- **What kind of backups should you perform?**
  - Partial: backing up only those files that have changed since the last backup
  - Full: backing-up all files
  - How often and what kind will depend upon what kind of data you have and how important it is
- **What about non-digital data (such as handwritten notes)?**
  - Consider digitizing files

Backup: 3-2-1 Rule

Best Practice:

- 3 Copies of datasets on
- 2 different devices/media of which
- 1 is off site



Where will you backup your files?

- Personal external disk (USB drive)
- **Network storage** (NAS, personal or institutional server)
- Cloud storage, e.g. Dropbox, Microsoft Onedrive, Google drive, iCloud, etc.

(May depend upon project requirements, etc.)

Backups: bottom line

- Pick a strategy
- Be consistent
- **Check if your data are restorable from the backups!**

Lesson 7: Data backup, security, storage & preservation

- ✓ Why?
- ✓ Where to store data
- ✓ Data backup
- → Data security
- → Data preservation

Information security

Threats & Defences

Undetected, unauthorized access

Backdoors, Exploits

Viruses, Trojans, Worms

Malware, Ransomware, spyware

Phishing

Logic bombs

Keylogging

etc.

Access control

Application security, e.g. Antivirus software

Authentication (e.g. multi-factor auth.)

Intrusion detection system

Encryption

Firewall

Access control

- **Are your data sensitive?**
- Who **has access** to the data?
- Who **is allowed to have access** to the data?
- How can data access be controlled? (Username / Password)

Image: Barry Levine, martechtoday.com @martech_today

Sensitive data: examples

- **Personal data:** identifiers such as names or identification numbers, physical, physiological, genetic, mental, economic, cultural or social characteristics.
  Also includes location data from GPS or mobile phones!

- **Confidential data**: trade secrets, investigations, data protected by intellectual property rights
  Security: passwords, financial information, national safety, military information...

- **Combination of different datasets** that can be combined into sensitive or personal data

- **Biological data**: endangered (plant or animal) species, where their survival is dependent on the protection of their location data.

- **Personal and sensitive metadata**

OpenAIRE Guide for Researchers:
How to deal with sensitive data
https://www.openaire.eu/sensitive-data-guide

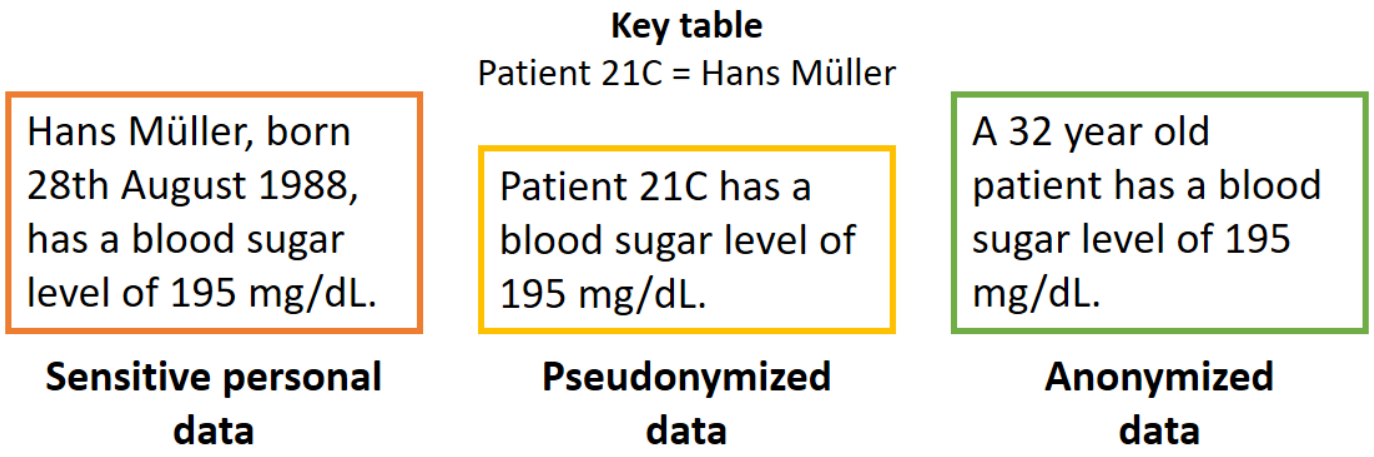
Data security measures for sensitive data

– **Anonymization: irreversibly destroy** any way of identifying the data subject.
Truly anonymized data is not sensitive anymore!

amnesia

amnesia.openaire.eu/
(data anonymization tool)

– **Pseudonymization:** substitutes the identity of a subject with a «nickname» (code, pseudonym). It is a secure approach if the personal identifiers are stored in a separate location.
Pseudonymized data are still sensitive data, because it can be linked back to the person!



Data security measures – not only for sensitive data!

Physical security

- Control access to buildings, rooms, cabinets where data, computers, media or hardcopy materials are held
- **No data on portable devices** (easily lost)

Network security

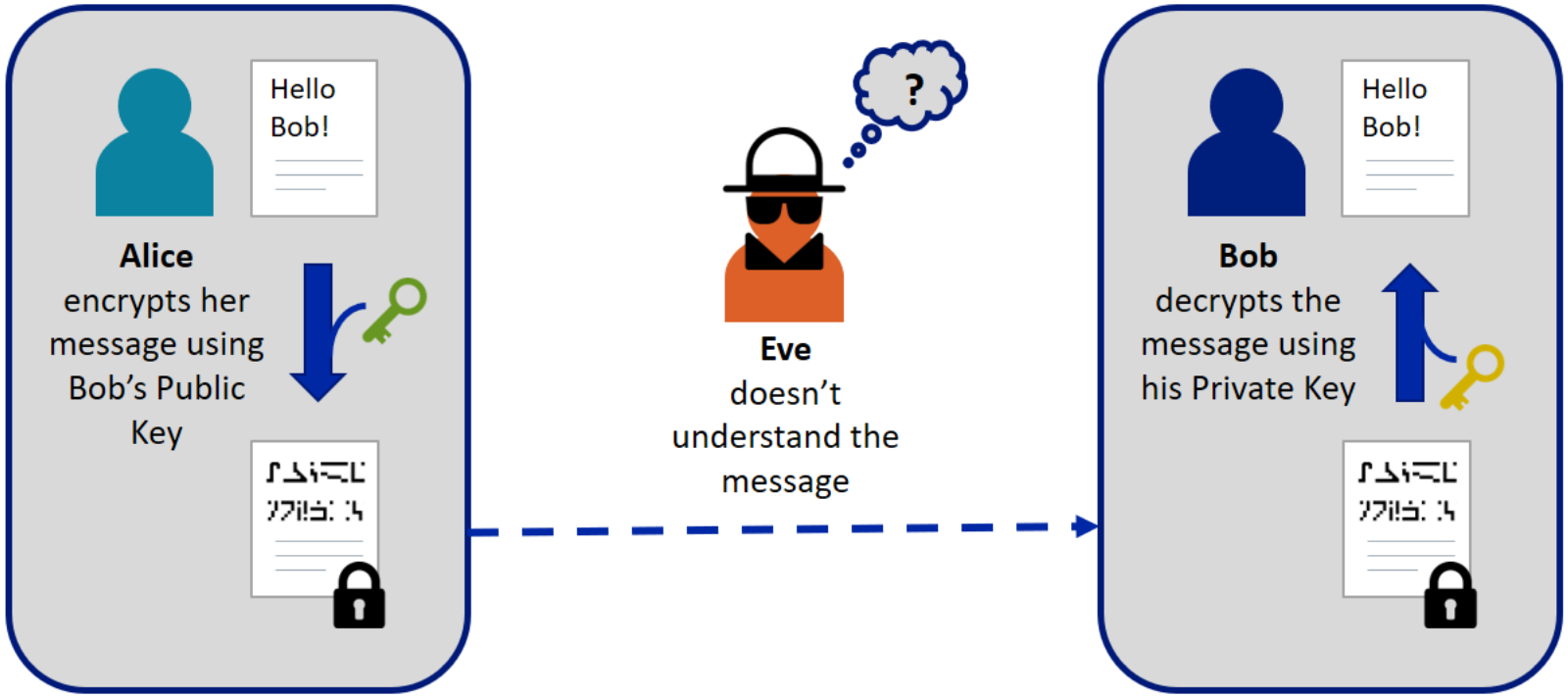
- Not storing sensitive data such as those containing personal information on servers or computers connected to an external network, particularly servers that host internet services
- **Firewall protection** and security-related **upgrades and patches to operating systems** to avoid viruses and malicious code

Security of computer system & files

- Lock computer with a password
- Ensure software is up to date.
- Implementing password protection and controlled access to data files, for example '**no access**', '**read only**', '**read and write**' or '**administrator-only**' permission
- **Not sending personal or confidential data via email or other file transfer** means without first encrypting them
- Special attention on **data disposal**! Formatting a hard drive will not prevent the possible recovery of data. More info: https://www.ukdataservice.ac.uk/manage-data/store/disposal
- Non-disclosure agreements for managers or users of confidential data.

Encryption

– Encryption is the process of encoding digital information in such a way that only authorised parties can view it.



Wikipedia article: Alice and Bob

– It is possible to encrypt individual files, folders, or entire disk volumes or USB storage devices.

– Encryption software: generates encryption / decryption keys (passwords).

Encryption Software

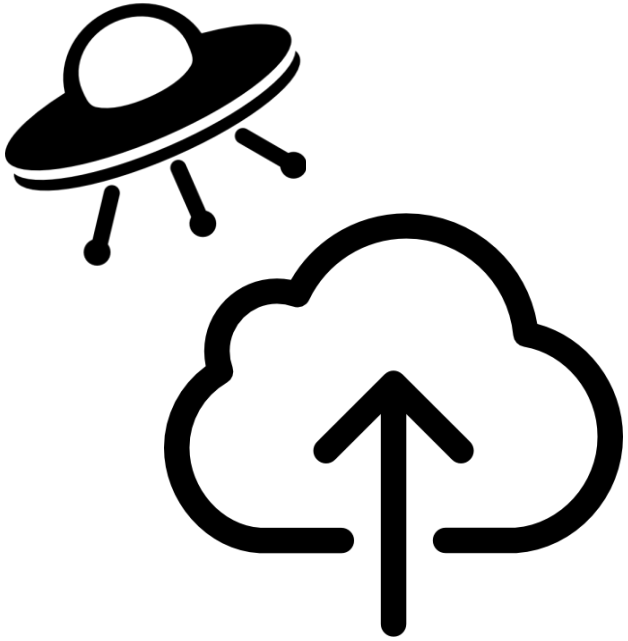
BitLocker
integrated in Windows (since Vista); offers encryption of disk volumes and USB devices

FileVault2
standard for Macs, (OS X Lion or later); full disc encryption

VeraCrypt
free, open source multi-platform encryption software (Windows, Mac and Linux);
full disk, partition and container encryption

Axcrypt
for Windows, Mac and mobile devices; basic version is free; premium version with monthly subscription and extensive features (e.g. secure files in Dropbox, Google Drive etc.).

Data security and cloud storage?



Before choosing a cloud storage service, ask yourself:

Where is the data actually stored when it is «in the cloud»?

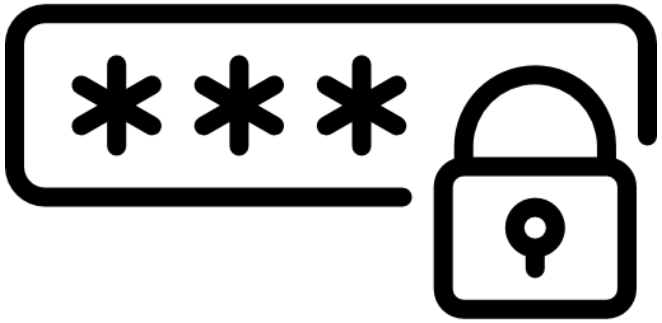
What kind of guarantees are there for data security?

- Servers should not be located outside of Europe
- Sensitive data should not go into the cloud!

- **Use SWITCH Drive!** SWITCH https://www.switch.ch/drive/
  - 100 GB storage
  - Synchronize files & folders
  - Desktop client, browser or mobile app
  - All data on Swiss servers and full compliance with Swiss data protection regulations.

Maybe the most important part…

Choose strong passwords!



1. Make your password long.
2. Make your password a nonsense phrase.
3. Include numbers, symbols, uppercase and lowercase letters.
4. No obvious personal information!
5. Do not reuse passwords!
6. Start using a password manager.
7. Don’t give your passwords away, do not write them down, etc.
8. Change passwords regularly

Information security and data protection at UZH

https://www.zi.uzh.ch/de/staff/it-security/information-security-and-data-protection.html

Scroll down for links to the laws of the canton of Zurich.

(unfortunately only in German..)

Exercise 7.3: Group Discussion

What are the pros and cons of portable storage devices?

Can you provide examples where researchers in your community have lost research data and/or when you have lost digital files yourself? How was the data lost and could it have been prevented?

What reasons might there be for researchers not to use networked drives to back up research data?

What are the principal reasons why researchers should back up their data files?

Do you use a cloud-based storage service?
Why or why not?

What methods could you use to manage sensitive data?

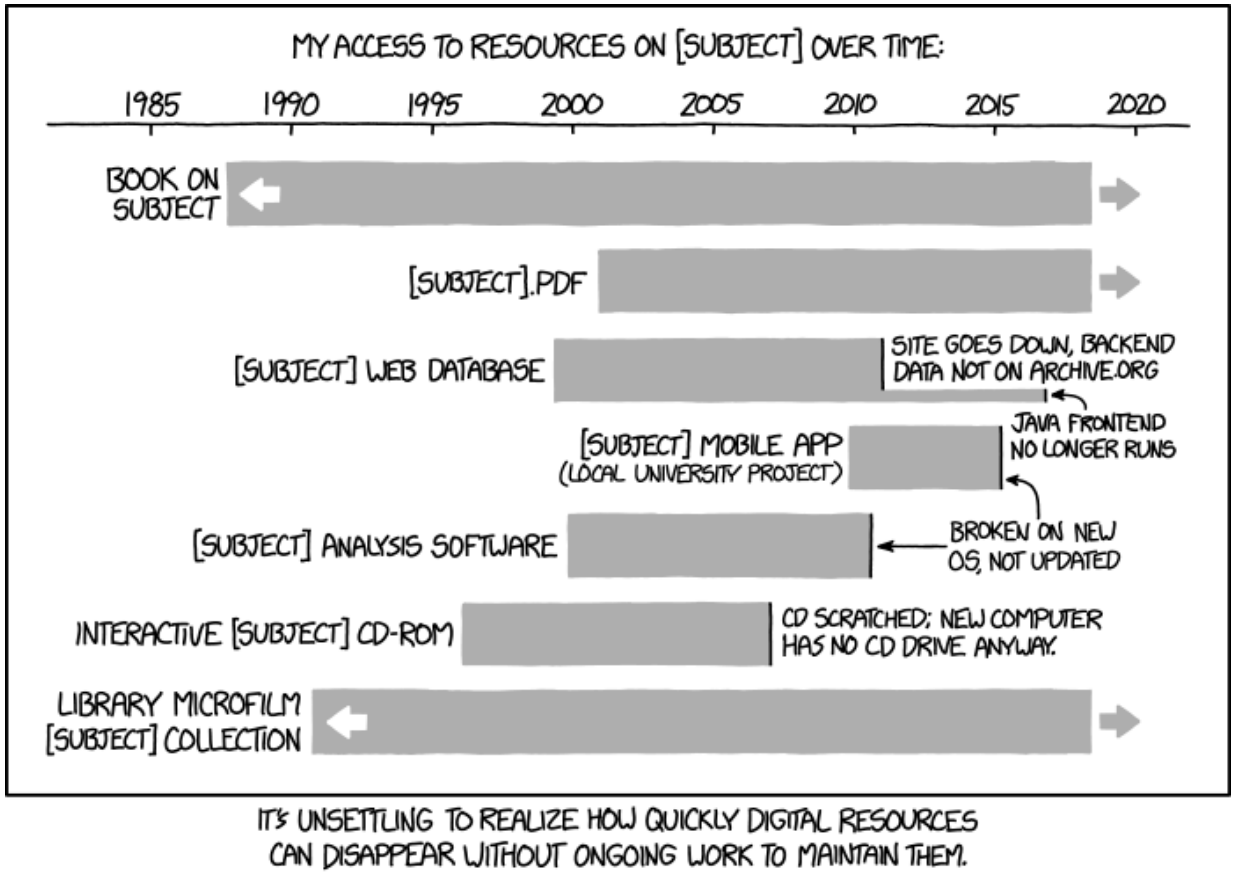
Lesson 7: Data backup, security, storage & preservation

- ✓ Why?
- ✓ Where to store data
- ✓ Data backup
- ✓ Data security
- → Data preservation

Digital long-term preservation

– Digital data feel like they are «everywhere» and «always here»

– But how long do they really last?



Questions for successful long-term management

- **What data will be kept** or destroyed after the end of the project?
  Make a selection.

- Are there requirements on **how long** your data needs to be preserved?
  (SNSF recommends at least 10 years)

- **Where will the data be archived?** Institutional server? Repository? Data archive?

- Are the chosen **file formats** long-lived?
  Programs and file formats change over time such that old files may become difficult to read.

- What **related information** will be deposited with the data?
  Metadata: if your data are not well described their preservation is useless.

Formats for archiving

Text

Recommended	Suitable only to a limited extent	Not suitable for archiving
• PDF/A (*.pdf) • Plain Text (*.txt, *.asc, *.c, *.h, *.cpp, *.m, *.py, *.r etc.) coded as ASCII, UTF-8, or UTF-16 using byte order mark • XML (inclusive XSD/XSL/XHTML etc.; with included or accessible schema and character encode explicitly specified)	• PDF (*.pdf) with embedded fonts • Plain text (*.txt, *.asc, *.c, *.h, *.cpp, *.m, *.py, *.r etc.) (ISO 8859-1 coded) • Rich Text Format (*.rtf) • Word *.docx • PowerPoint *.pptx • LaTeX, TeX OpenDocument formats (*.odm, *.odt, *.odg, *.odc, *.odf)	• Word *.doc • PowerPoint *.ppt

Formats for archiving

Spreadsheet or table

Recommended	Suitable only to a limited extent	Not suitable for archiving
• Comma- or tab delimited text files (*.csv)	• Excel *.xlsx • OpenDocument spreadsheets (*.ods)	• Excel *.xls, *.xlsb

Formats for archiving

Images and Graphics

Recommended	Suitable only to a limited extent	Not suitable for archiving
• **TIFF** (*.tif) (uncompressed, preferentially TIFF 6.0, Part 1: baseline TIFF). • Portable Network Graphics (*.png, uncompressed) • JPEG2000 (*.jp2, lossless compression) • SVG without JavaScript binding (*.svg)	• TIFF (*.tif) (compressed) • GIF (*.gif) • BMP (*.bmp) • JPEG/JFIF (*.jpg) • JPEG2000 (lossy compression) (*.jp2)	• InDesign (*.indd), • Illustrator (*.ait) • Encapsulated Postscript (*.eps) • Photoshop (*.psd)

Formats for archiving

Audio and Video

Recommended	Suitable only to a limited extent	Not suitable for archiving
• WAV (*.wav) (uncompressed, pulse-code modulated) • FFV1 codec (version 3 or later) in Matroska container (*.mkv)	• Advanced Audio Coding (*.mp4) • MP3 (*.mp3) • MPEG-2 (*.mpg,*.mpeg) • MP4, which is also called MPEG-4 Part 14 (*.mp4) • QuickTime Movie (*.mov) 2) • Audio Video Interleave (*.avi) • Motion JPEG 2000 (*.mj2, *.mjp2)	• Windows Media Video (*.wmv)

Summary of Lesson 7

Choose the right storage medium for your data.

Recognize sensitive data and treat them accordingly.

Don’t be lazy with backups! Remember the 3-2-1 rule.

Regularly assess the security of your data. Consider using an encryption software.

Choose strong passwords and change them regularly to keep your data secure.

Think of the long-term preservation of your data. Curate your data, choose appropriate file formats for archiving and create quality metadata.