# Data Management:
# an introduction

# Data management is more than storage

- Before we can talk about storage, we have to talk about why we store data
- Will your group's data be usable in:
  - 5 years – all your group has changed
  - 50 years – all department/university servers have changed?
- Are you already suffering from mistakes of the past?
- People who fund research are starting to insist that data not be lost.

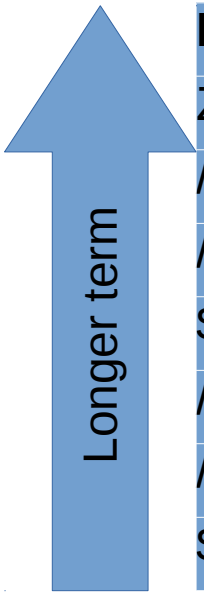
# Types of data

Key point of data management: there is different types of data.  Keep it organized.

- Original data – keep long time/forever
- Code – keep forever
- Scratch intermediate files – temporary
- Final results – keep forever
- Published data/code/paper – keep forever

# Storage locations

- Many places to store data already
- This example for CS, but applies to other departments too




| Location | Use |
|---|---|
| Zenodo / EUDAT | Permanent EU-level archive |
| /m/cs/archive | Medium-term storage |
| /m/cs/project | Normal backed-up project files |
| $HOME | Personal files, not shareable |
| /scratch and /work | Large scratch non-backed-up space |
| /tmp (local disk) | Small and random access |
| $XDG_HOME_DIR (ramdisk) | Extremely temporary |

# Directory organization

- Keep your files organized.  I often used:

  original/

  code/

  scratch/

  doc/

- Multi-user collaboration:

  original/

  user1/{code,scratch,doc}/

  user2/{code,scratch,doc}/

-

# Mary other considerations

| | **Short-term** | **Long-term** |
|---|---|---|
| **Technical** | Sharing<br>Security<br>Backups<br>Organization<br>Version control | Physical storage locations<br>Archiving<br>File formats<br>Clean-up |
| **Non-technical** | Documenting<br>Collaboration<br>Common standards<br>IP rights<br>Ethical evaluation | Data repositories<br>Citation<br>Publishing<br>Ownerships<br>Licensing |

# Summary

- Think before you act
- Your entire group will be different in five years: will anything you do now be usable to them?
- Ask if you have questions: colleagues and staff