

Data management

Andy J. Wills

Keeping organized

- Critical to keep organized records of:
 - what was done
 - by whom
 - When
- Assessment requirement: Share records with supervisor
- Open science: On publication, sharing anonymized data, materials, analysis scripts publicly

github

- Using **github** repositories as your data archive and log book.
- **Github:**
 - 40 million users
 - Acquired by Microsoft in 2018 for \$7.5bn
 - Based on git, the system used to develop Linux

Activities today

- Create a private github repository
- Sharing with your team
- Accessing from Rstudio
- Adding a file
- Modifying a file
- *git log* as your lab logbook
- Branching, and file recovery

Anonymity and privacy

- Your repository may not always be as private as you think
- Open science sharing
- **Never put these on github:** names, addresses.
- **Explicit supervisor approval for:** biographical information (age, gender, education, etc.). Risk of *de-anonymization*
- **Never make your repo public** (leave that to your supervisor).
- **Tell supervisor** immediately if you've added anything potentially de-anonymizing (even if you remove it again immediately, see later)

Good and bad file types

- **Best:** plain text e.g CSV, R, TXT, markdown
- **Sometimes unavoidable:** Picture files (JPG, PNG)
- **If you have to:** Wordprocessor files: use Open Document Format (e.g. LibreOffice) rather than Microsoft Office format.
- **Never:** Excel, SPSS.

Assessment

- Set up a private repository for your dissertation
- Add at least one relevant file
- Use github to share it with me.
- Once you've done all that, screenshot the front page of your github repository on <https://github.com>, convert to PDF, and upload to Psyc:EL.