INTRO TO REPRODUCIBLE RESEARCH

SNAP workshop
Feb 20th 2025

daniel.wrench@vuw.ac.nz

WHAT IS SNAP?

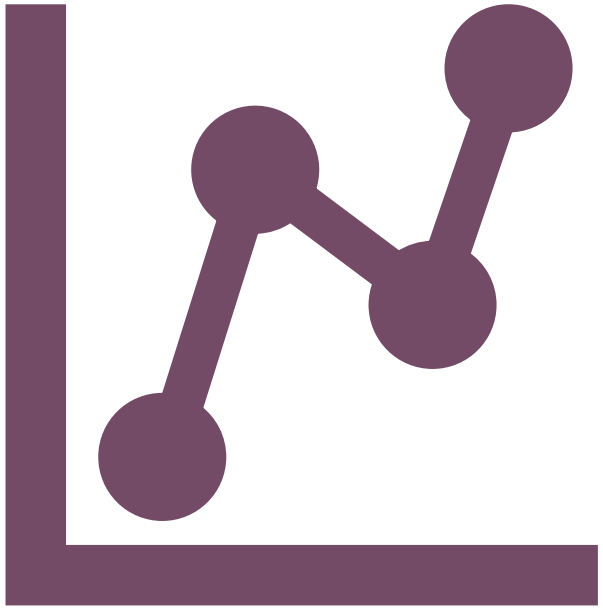
SNAP

VUW HUB FOR SIMULATION, NUMERICAL METHODS, ANALYTICS AND PROGRAMMING

- Interdisciplinary community of researchers who code, simulate, use supercomputers (e.g. Rāpoi), etc.
- Sharing expertise across subjects
- Looking for student rep
  - Helping connect with postgrad student community
  - Monthly meetings

SUMMARY OF WORKSHOP

- **What is reproducible research?**
- **Creating a reproducible data pipeline**, as one would regularly encounter in scientific analysis or data science
  - **Good code repository structure**
  - **Using Git and GitHub**
  - **Virtual environments**
  - **Sharing your codes**
- Won't cover:
  - Containers
  - Pull requests and other intermediate/advanced aspects of Git
  - Object-oriented programming
  - Testing, `__init__.py` files, and other aspects of creating a piece of "software"

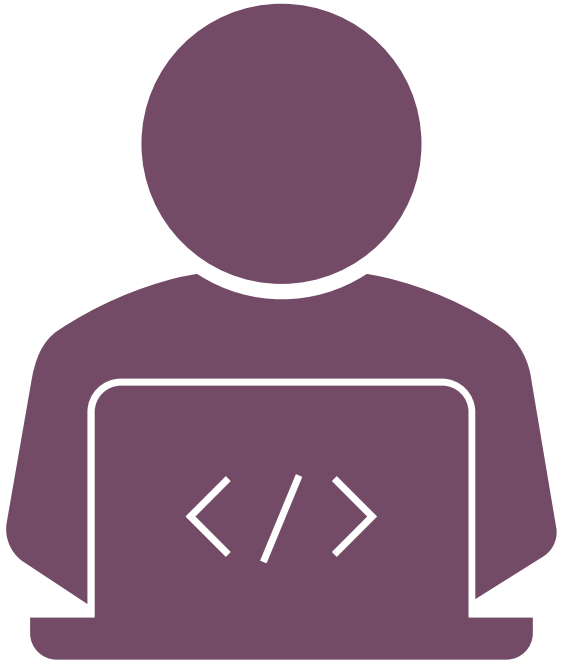


YOU WILL NEED

- The following programs installed
  - Git (can provide link to step-by-step instructions)
  - Python
- An account on GitHub.com
- A terminal or IDE of your choice (I'll be working in the terminal and VS Code)
- **Ask if you need help getting set up**

WHAT IS REPRODUCIBLE RESEARCH?

- AKA open science, sustainable research software
- **What do these terms mean to you?**
- Set of principles and practises in programming and documentation that ensure
  - Collaboration
  - Longevity
  - Transparency
- 3 key pillars
  - Project structure
  - Version control
  - Virtual environments



TASK 1: **GET STARTED WITH A REPO TEMPLATE**

1. Go to https://github.com/daniel-wrench/snap-research-template
2. **Understand Clone vs. Fork vs. Use template**
3. Use this template -> Create a new repository
4. Give it your own name
5. Wonder at your perfectly-structured creation!
6. Code -> Local -> copy HTTPs URL
7. Open a terminal (in VS Code, Git Bash, whatever) and navigate to where you want to work
8. `git clone paste_url_here`
9. `cd your-repo-name`

OR START WITH AN EXISTING FOLDER ON YOUR COMPUTER

```
git init
git remote add origin github_url.git
```

TASK 2: **SET UP A VIRTUAL ENVIRONMENT**

1. **What is a virtual environment?**
2. Follow steps in README
3. In step 3, first `pip install` the following packages
   - `pandas`
   - `requests`
   - `matplotlib`
4. Make your first commit! (requirements.txt)
   a. `git status`
   b. `git add requirements.txt`
   c. `git commit -m “informative-commit-mssg”`
   d. `git push`
   e. `git status`

TASK 3: **DOWNLOAD DATA**

1. Run the code in `scripts/` from the terminal or VS Code play button
2. **Can/should we commit this file?**
3. Delete from the terminal
4. Correct the `output_path` in the script
5. Re-run
6. Note `git status`
7. **What if I only wanted to ignore the raw data, not the processed stuff?**
8. Pull up changes made to the script
9. Commit this change

TASK 4: **PLOT DATA**

1. **Make a plot of this data.** Up to you how to do it: doesn't need to be anything fancy. Think about where the code should go, a good name for it, etc.
2. Commit this file
3. Delete a bunch of the file and save it
4. Check with `git status`
5. Undo the change: `git restore file_name`
6. Change it again, commit, push
7. Undo the commit: `git revert HEAD`

DEMO OF USING GIT WITH HPC CLUSTER

RĀPOI

VUW's High Performance Compute Cluster

NeSI

New Zealand eScience Infrastructure

SHARING YOUR REPO

- **Want reproducible analysis not just for ourselves or our colleagues, but for the whole scientific community.** Share data and software in your papers!
- **How could we do this?**
- All too common: *The data (and maybe poorly documented code) are available on reasonable request*
- Better: *Here's the link to my GitHub repo*
- Best: *The code is available on GitHub (link) and* ***is archived in Zenodo (citation with DOI)***
- Zenodo-GitHub integration -> CITATION.cff file in repo -> easy copy-and-paste BibTex citation
- **Finally, ensure you have good documentation for when they get the codes!** At minimum, a comprehensive README and metadata (explanation of the data and where it came from)

FINAL THOUGHTS

- Feel free to use my template however much you want – just remember the "Use this template" button
- This is all extra work, but it's worth it: for you, for your colleagues, and for science
- It's also not the whole picture: need tidy, readable, documented, modular code as well!
- ChatGPT and other LLMs are an invaluable tool – as long as you're not blindly copy-pasting!
  - For VS Code users, highly recommend installing Copilot. Limited version free to everyone, unlimited if you sign up to a GitHub student account
- Any volunteers for SNAP student rep?

ANY QUESTIONS?

INTRO TO REPRODUCIBLE RESEARCH

SNAP workshop
Feb 20th 2025

daniel.wrench@vuw.ac.nz