UK Data Service

Reproducibility Collaborative Working

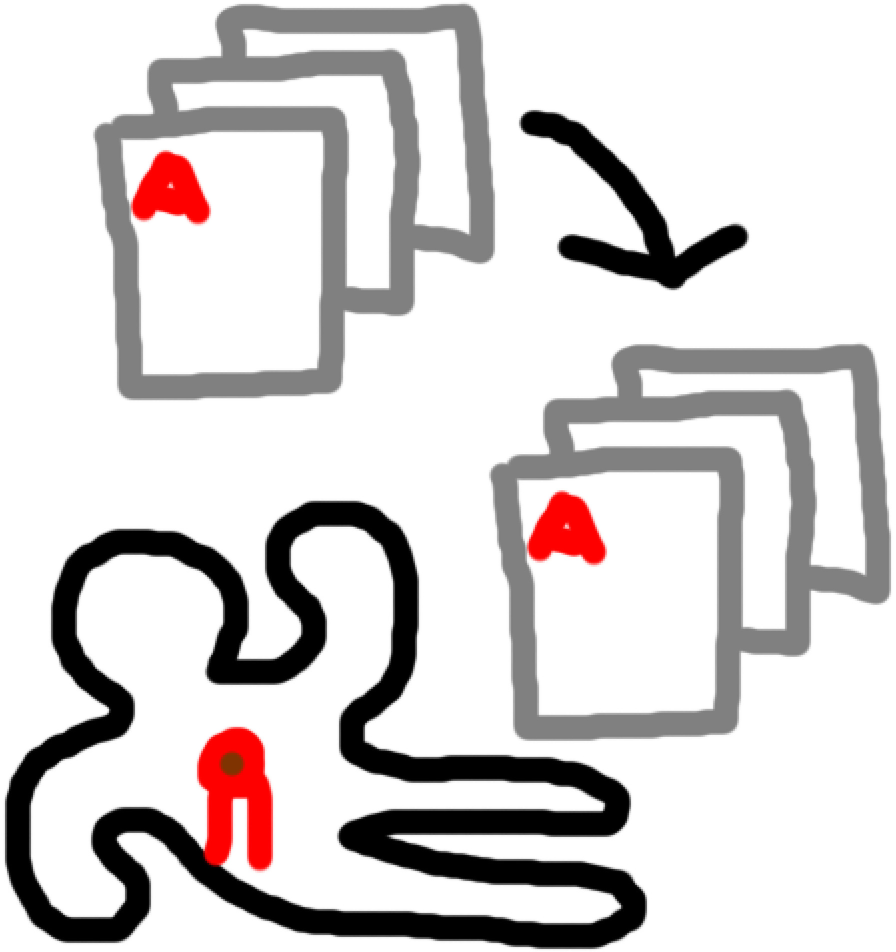
Joseph Allen – Research Associate at the UK Data Service

Joseph.Allen@Manchester.ac.uk

UKRI Economic and Social Research Council

In this webinar

- The problems of research
- Writing an English essay
- What is reproducibility?
- What is collaborative working?
- What is version control?
- What are git and GitHub?
- Git terminology
- A crime scene case study
- GitHub Demo



Aims

- Explain what git is
- Explain what GitHub is
- Clone a repository
- Add files to a repository
- Edit files on a repository

Workshop warning

If you want to follow along for the GitHub demo you will need:

- A GitHub account - github.com
- GitHub Desktop downloaded and set up - desktop.github.com
- To be logged into GitHub Desktop with your GitHub account

The problems of working with data in research.

What makes research hard?

- Useful data is protected
- Methods are protected
- We encourage positive results
- In some fields, most results are not reproducible
- Academic work is hard
- Accountability is boring

Useful data is protected

- Data is protected, and should be.
- Data is dangerous.
- Without protection, there would be limited access.
- Access should be given – research and verification.
- The future – data access given to academics and journals purely to verify results.

Methods are protected

- Fear of “stolen” work.
- Delay methods until publication.
- Loss of unpublished methods.
- Academics forget what they did.

We encourage positive results

- We publish more positive results
- Controversial results get adopted by media
- Academics could tweak analysis to force a result

Most results aren’t reproducible

- Methods unavailable
- Data unavailable
- “70% of researchers have tried and failed to reproduce another scientist’s experiments” (Chambers 2014)

Academic work is hard

- Academics are the experts of their field
- Trying to do something nobody has ever done
- Little time for extras

Accountability is boring

- Timesheets may be skippable
- Staff are too busy to check accountability work
- Why add extra, boring work?

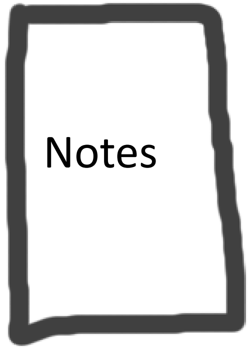
The bar is low

How is freedom represented in the Shawshank Redemption?

How would you write an essay?

- 5000 words on “How is freedom represented in Shawshank Redemption?”
- Consider the following:
  - How would you keep notes?
  - How do you manage your drafts?
  - Do you get feedback?
  - How do you manage your feedback?

The Notes stage

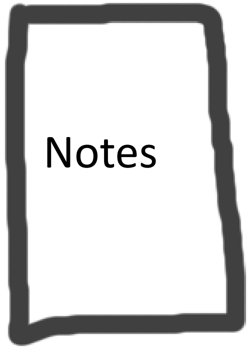


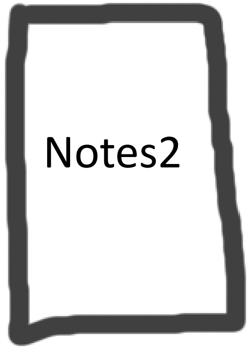
Notes

- To start with I might break down the entire essay:
  - Introduction
  - Point 1 – The use of bird imagery
  - Point 2 – Progression of time
  - Conclusion
- Save as – Notes.txt

Enhanced notes

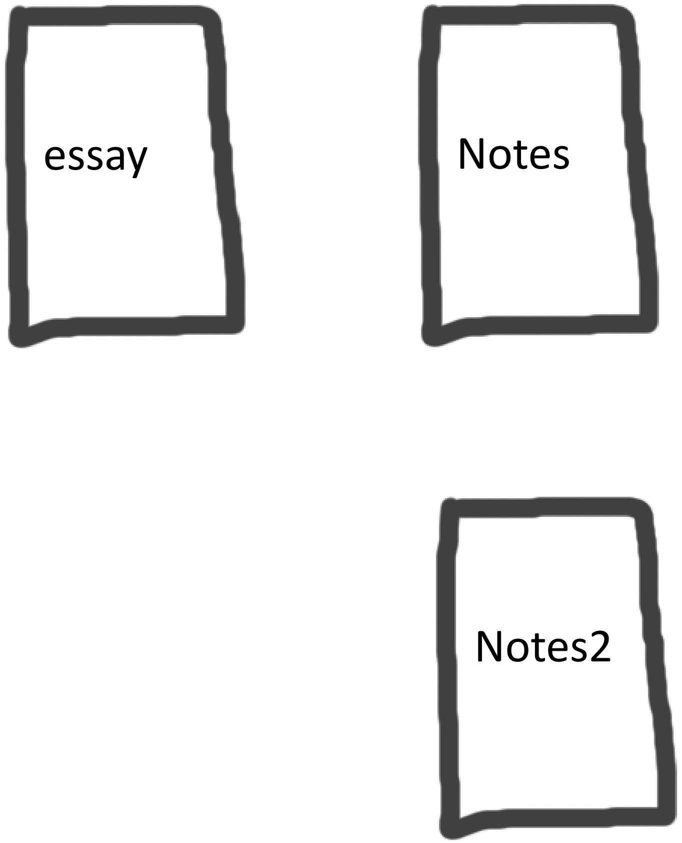
- Introduction
  - Introduce the question
- Point 1 – The use of bird imagery
  - Pet crows
  - The use of birds eye shots
- Change Point 2 – The rock hammer
  - Foreshadowed by Red
- Conclusion
  - Freedom = escape?
  - Bird in a cage?
- Save as – Notes2.txt





Flesh out points

- Point 1 – The use of bird imagery
  - The pet crows found in the prison, unable to fly represent Andy and the other prisoners.
  - Point 2 – The rock hammer
    - Red foreshadows the use of the rock hammer, saying it would take ”100 years” for somebody to break out of a prison with it.
- Save as – essay.txt – little indication of the current state of the project.



Write a conclusion

- Conclusion
  - In many cultures a bird represents freedom. A dove represents peace. Until recent history flight was not something humanity could achieve, now only enabled by technology inaccessible to many. The juxtaposition of a bird, ready to fly anywhere, capture in a cage is also a common icon.
  - The Shawshank redemption frequently reminds us of the icon of flying away. The prison yard is open air, a problem to Andy but not to his pet crow. Frequent birds eye shots are used to give the viewer a sense of that freedom, whilst reminding us Andy is trapped.

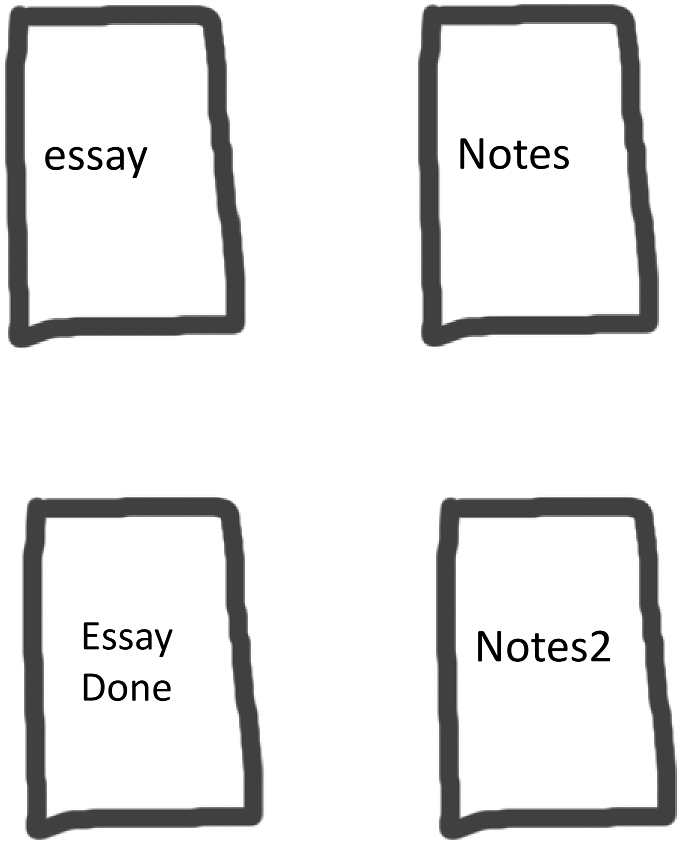
First draft

- I add an Introduction
    - Introduction
        - What is freedom but a limit we have not conquered. Does the likelihood of this limit being overcome affect our upset with the limit? In this essay I seek to answer how freedom is represented in the movie The Shawshank Redemption.

Powercut…

First draft - again

- I add an Introduction… again.
- I add a conclusion … again.
- Save as – essayDone.txt



essay

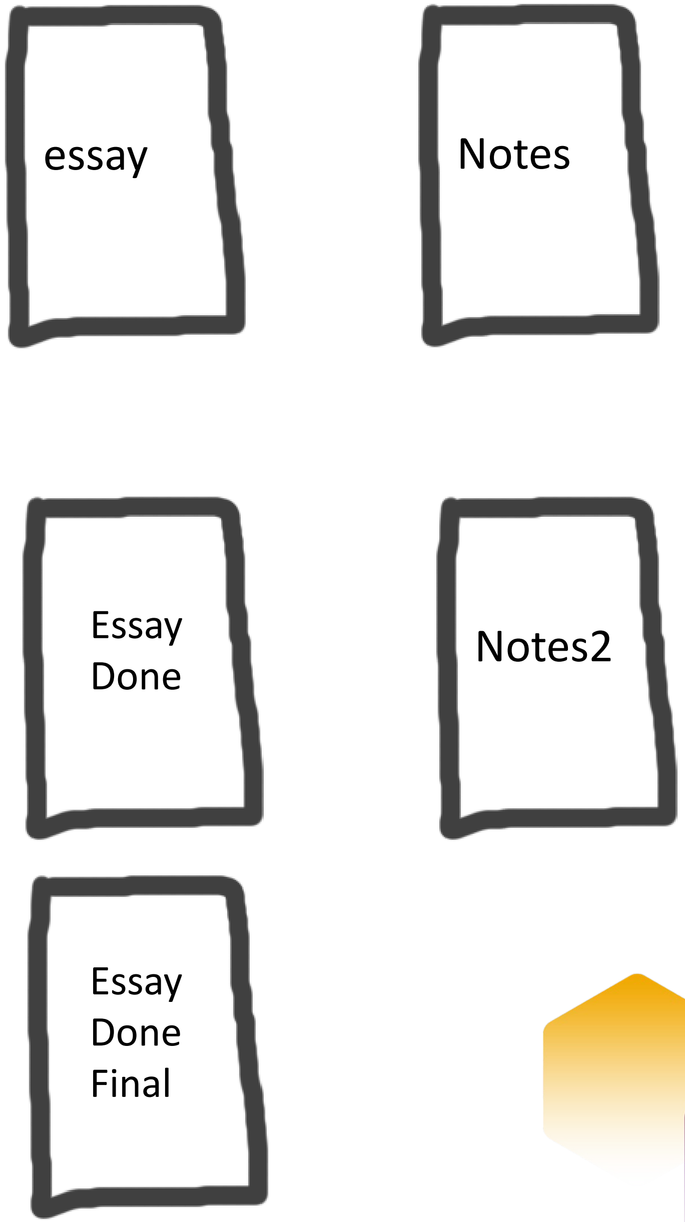
Notes

Essay Done

Notes2

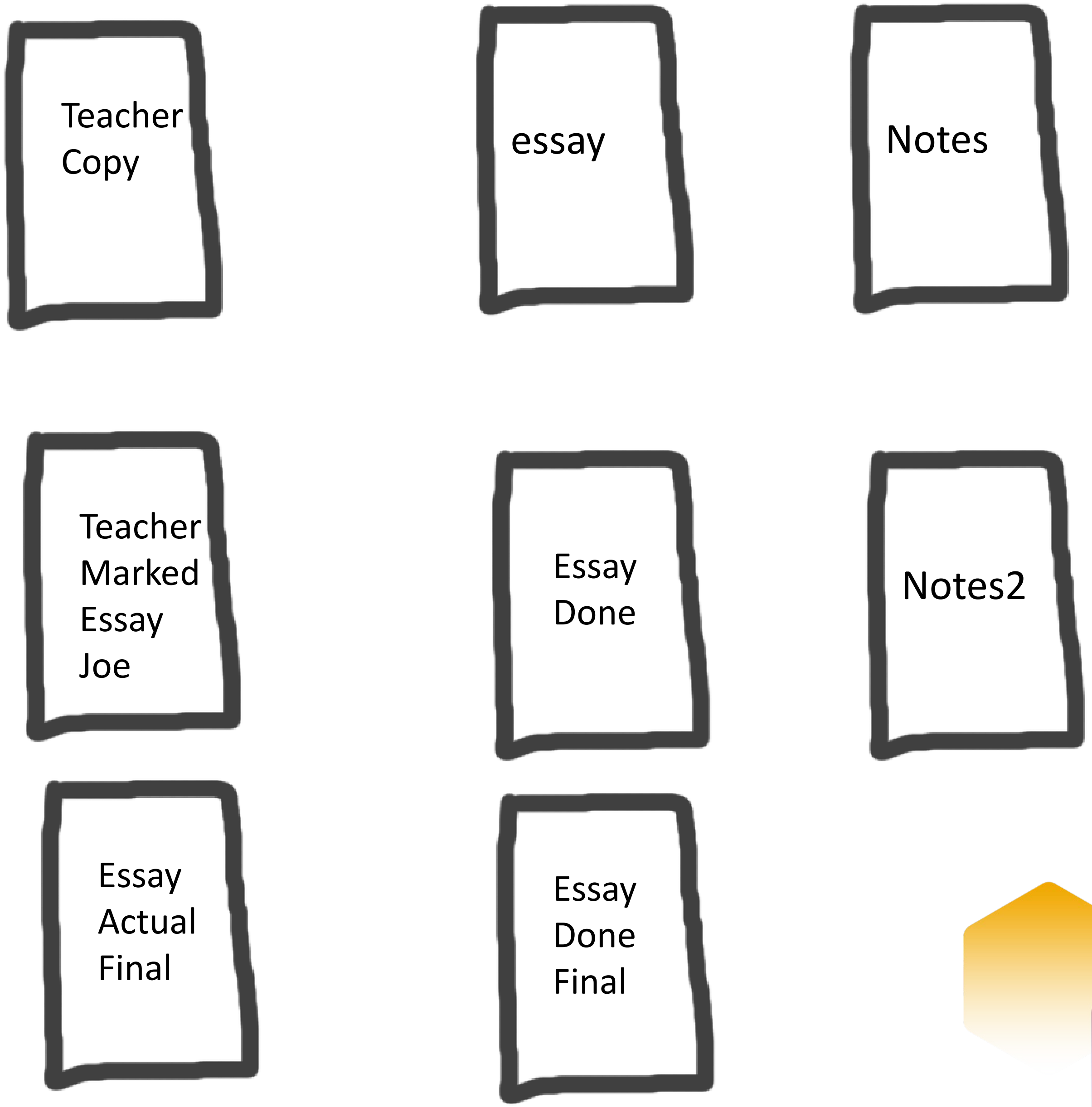
Spellcheck

- I proof read, fix any obvious spelling issues and more.
- Save as – EssayDoneFinal.txt, no indication of what I fixed, or what is yet to be fixed.



Review

- My teacher kindly offered to review my work.



Teacher Copy

essay

Notes

Teacher Marked Essay Joe

Essay Done

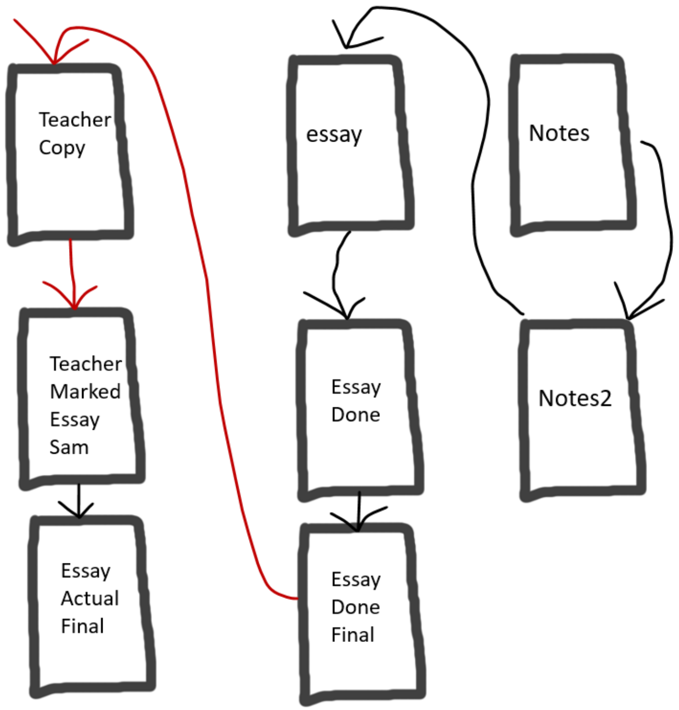
Notes2

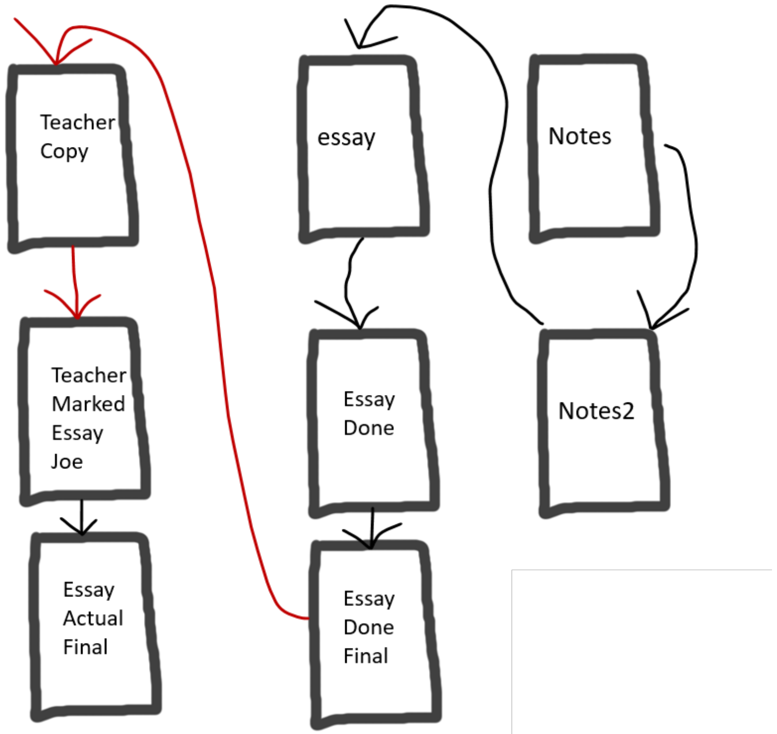
Essay Actual Final

Essay Done Final

My twin brother

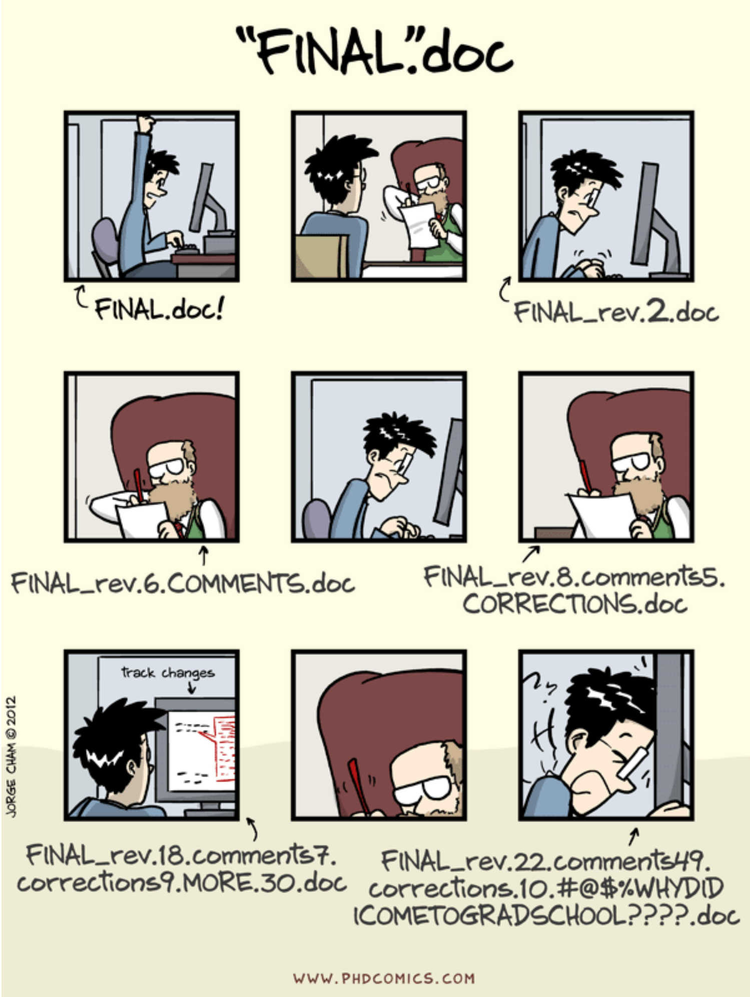
- My twin brother - same class, same task, same computer
- My brother accidentally submits essayActualFinal.doc
- Both submitted the same thing





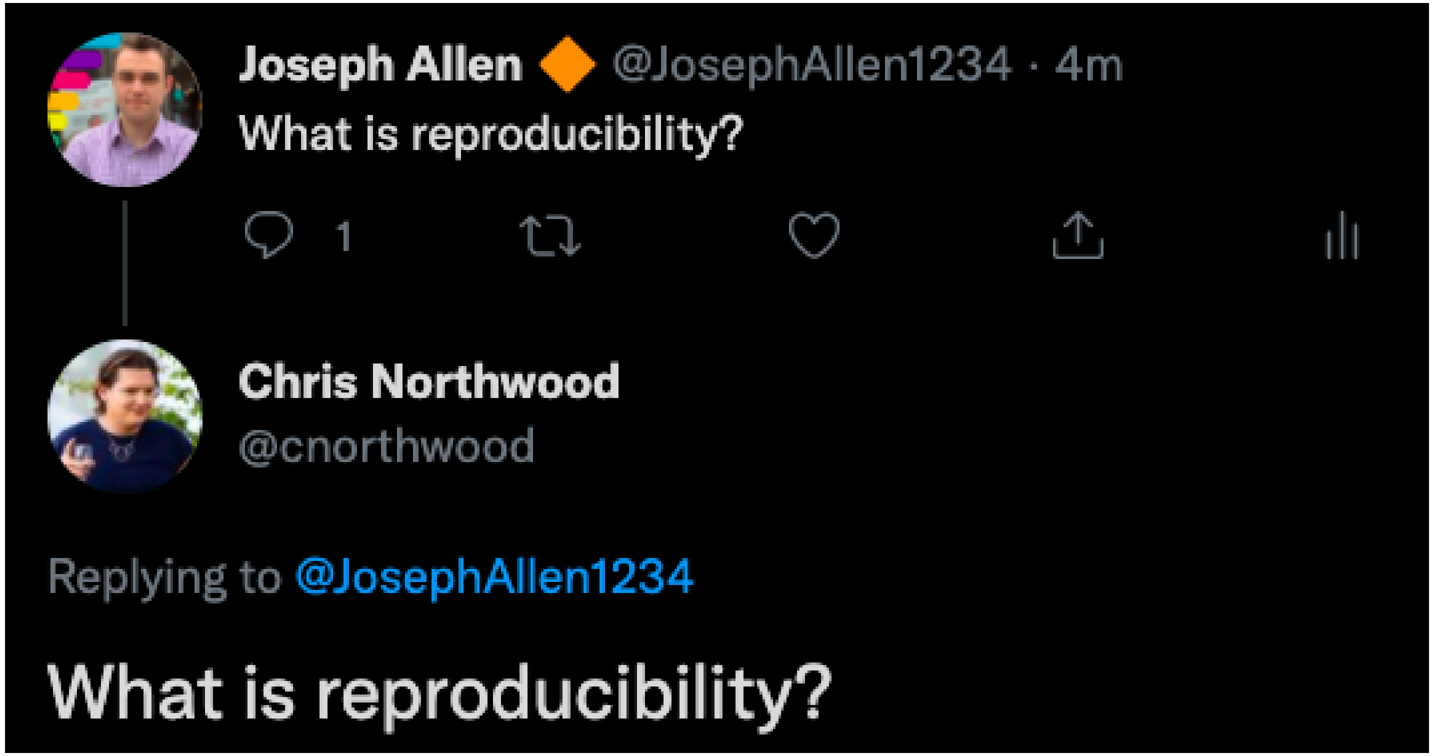
What went wrong?

- Overwritten files
- Context hidden in filenames
- Deletion of work
- Single fail point
- Collaboration creates complexity
- No accountability we did any work


"FINAL".doc
FINAL.doc!
FINAL_rev.2.doc
FINAL_rev.6.COMMENTS.doc
FINAL_rev.8.comments5. CORRECTIONS.doc
track changes
FINAL_rev.18.comments7. corrections9.MORE.30.doc
FINAL_rev.22.comments49. corrections.10.#@$%WHYDID ICOMETOGRADSCHOOL????.doc
JORGE CHAM © 2012
WWW.PHDCOMICS.COM

The solution - Reproducibility

What is reproducibility?



Joseph Allen @JosephAllen1234 · 4m

What is reproducibility?

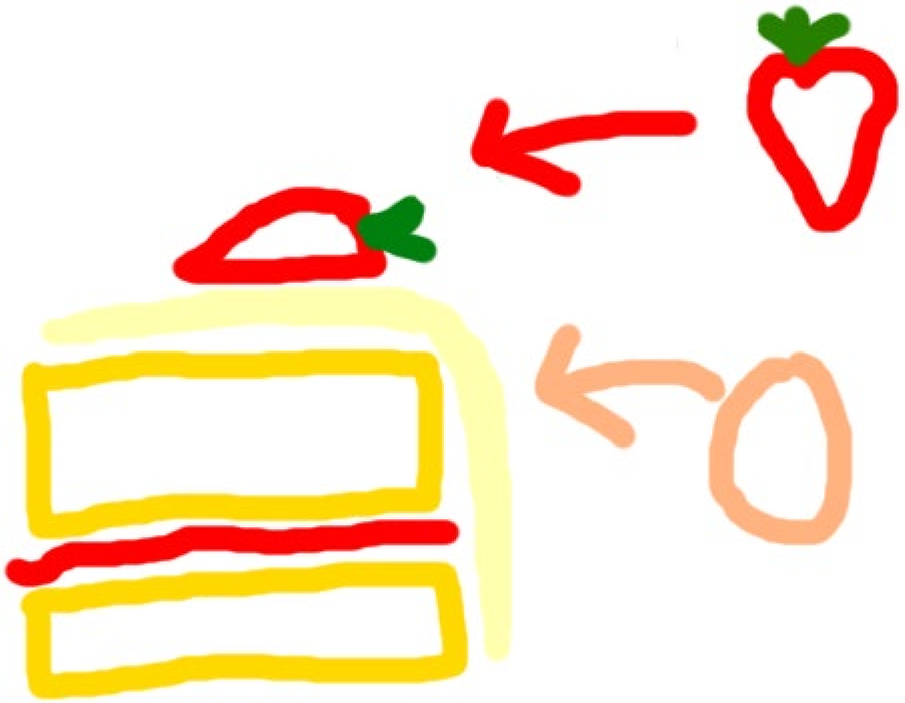
Chris Northwood
@cnorthwood

Replying to @JosephAllen1234

What is reproducibility?

How to make work reproducible

- Data – Acquisition, synthetics and more
- Tools – materials, software etc.
- Decisions made
- Results
- Access – if possible
- Bonus – Proof of work

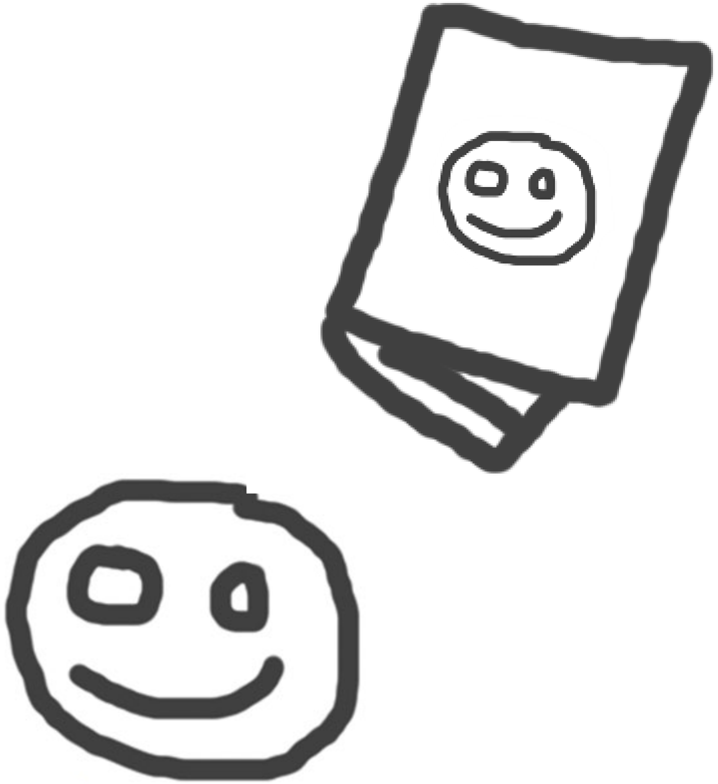


The author benefits

- Don’t forget why you did something
- Analysis is well documented
- Easy writing process
- Verifiable work
- Proof it’s your work

The journals benefit

- Lower risk of an academic ”scandal”.
- Journals can set a higher standard
- Verification of results is simpler
- Negative results are more valuable.

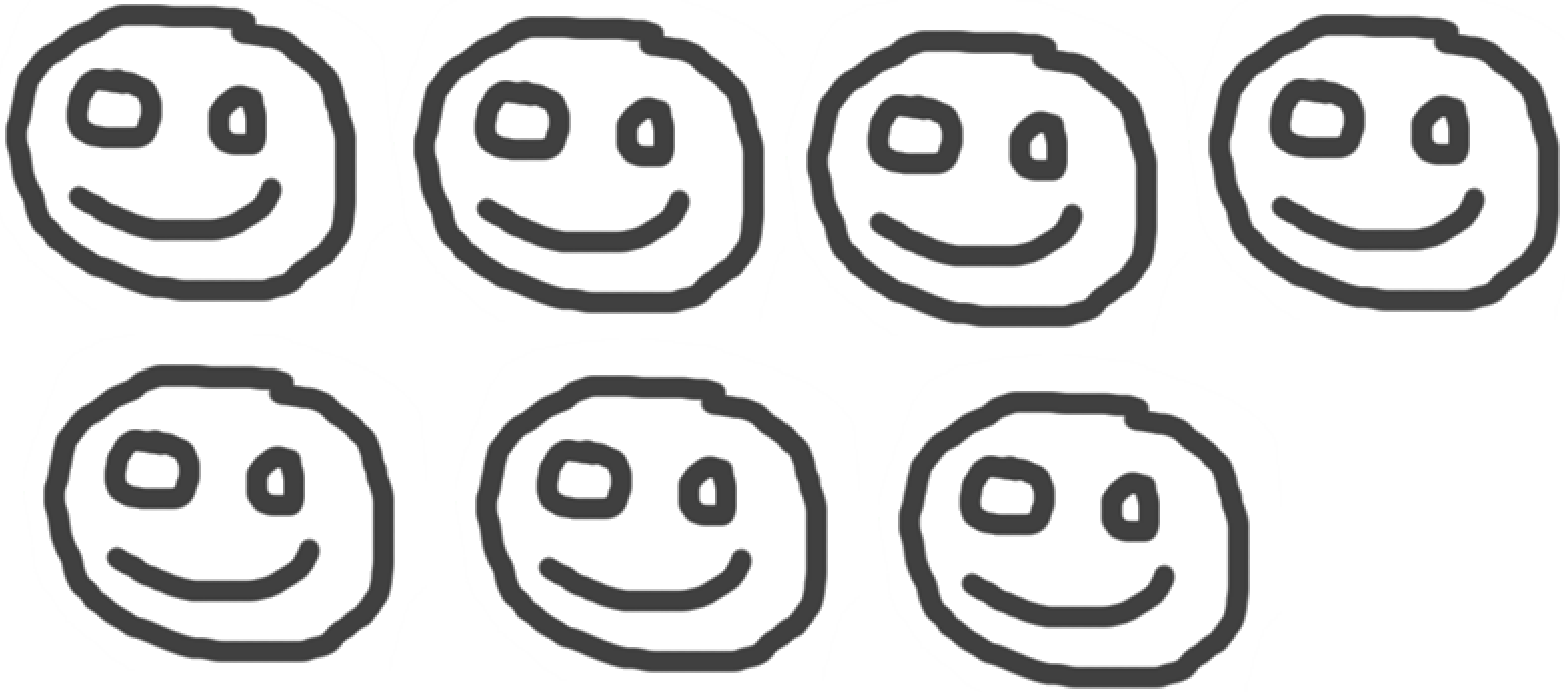


Acaedmics benefit

- Data access is well-documented
- Replicating methods is trivial
- More time freed for reproducibility in the next generation
- Reduces repeat work

The general public benefits

- Data professionals verify and enhance work
- Government policy more accountable



What is collaborative working?

“Working with somebody to produce something”

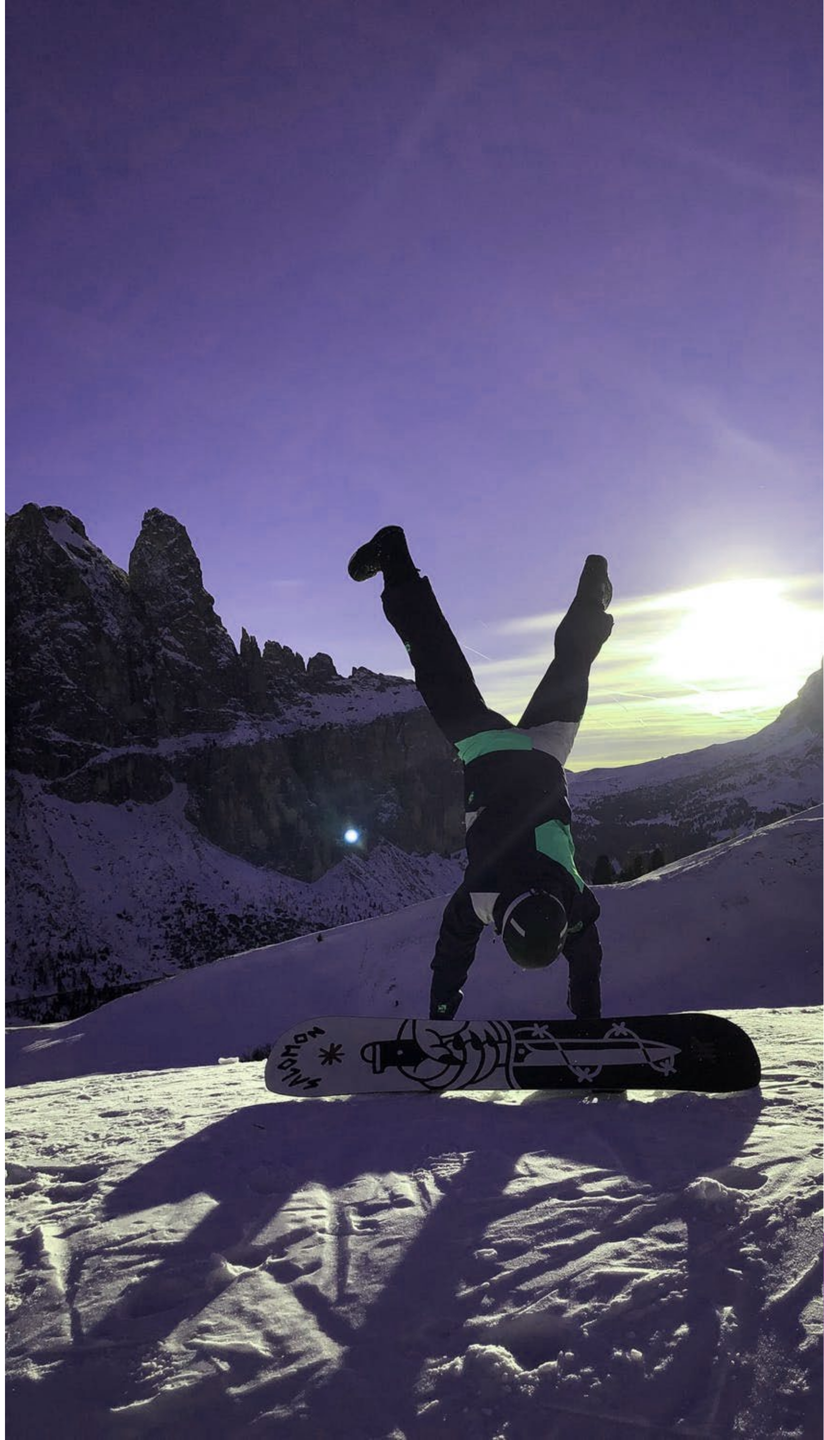
What is collaborative working?

- The person most likely to want to reproduce our work, is ourselves.
- Collaboration can be working with ourselves
- Sharing results, methods, making results and methods open
- Share the load – sometimes we need to work with other people who have different skillsets.

“Collaboration is the intention for your work to be built on by anyone”

The problems get easier

- Useful data is protected, but accessible
- Methods are shared
- We encourage results
- Future results become easier to reproduce
- Academic work is still hard
- Accountability is still boring, but quick



What is version control?
What is git?

What is a Version Control System(VCS)

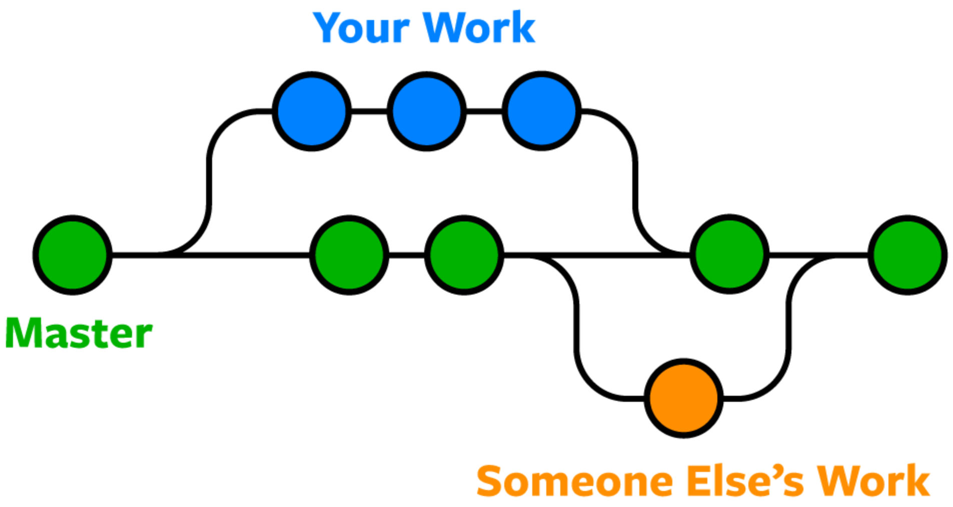
- A piece of software which allows you to record and preserve the history of changes made to directories and files.

Why use a VCS?

- Files comprise one project
  - The current version is on your computer
  - All previous versions are saved in a repository
- All changes are stored
- Commit messages describe changes
- Changes can be reverted
- Collaboration
- Responsibility

What is git?

- Git is a VCS
- Was built to help develop the Linux operating system
- Usually a command line tool



What is a repository?

What is a repository?

- Storage of a collection of files
- Sometimes called a "repo"
- A repository may be:
  - Locally hosted
  - An Online service like GitHub

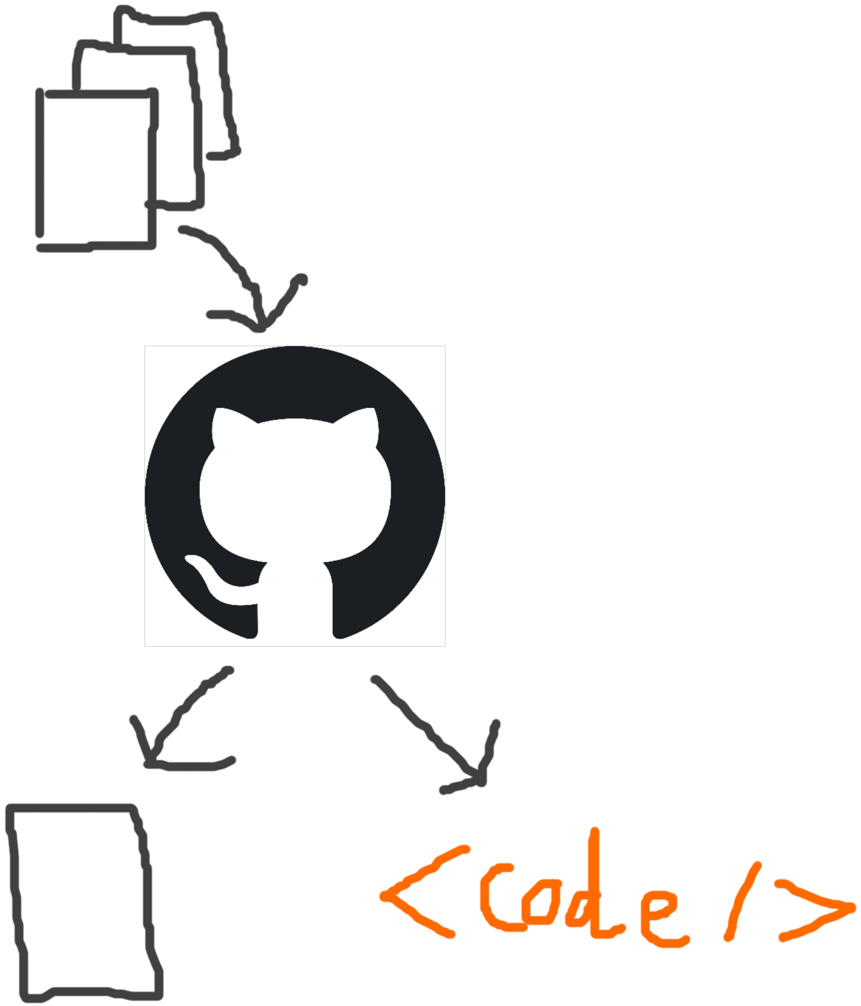
What is GitHub?

What is GitHub?

- GitHub is a place for files to live not on our computer.
- GitHub is a social media platform
- GitHub works with git

Useful features of GitHub

- Every project has README
- GitHub renders:
  - PDF and similar document files
  - Code with syntax highlighting
- Services know to accept GitHub links
  - Binder will deploy notebooks
  - Zenado creates citable DOIs
- Facilitates collaborations



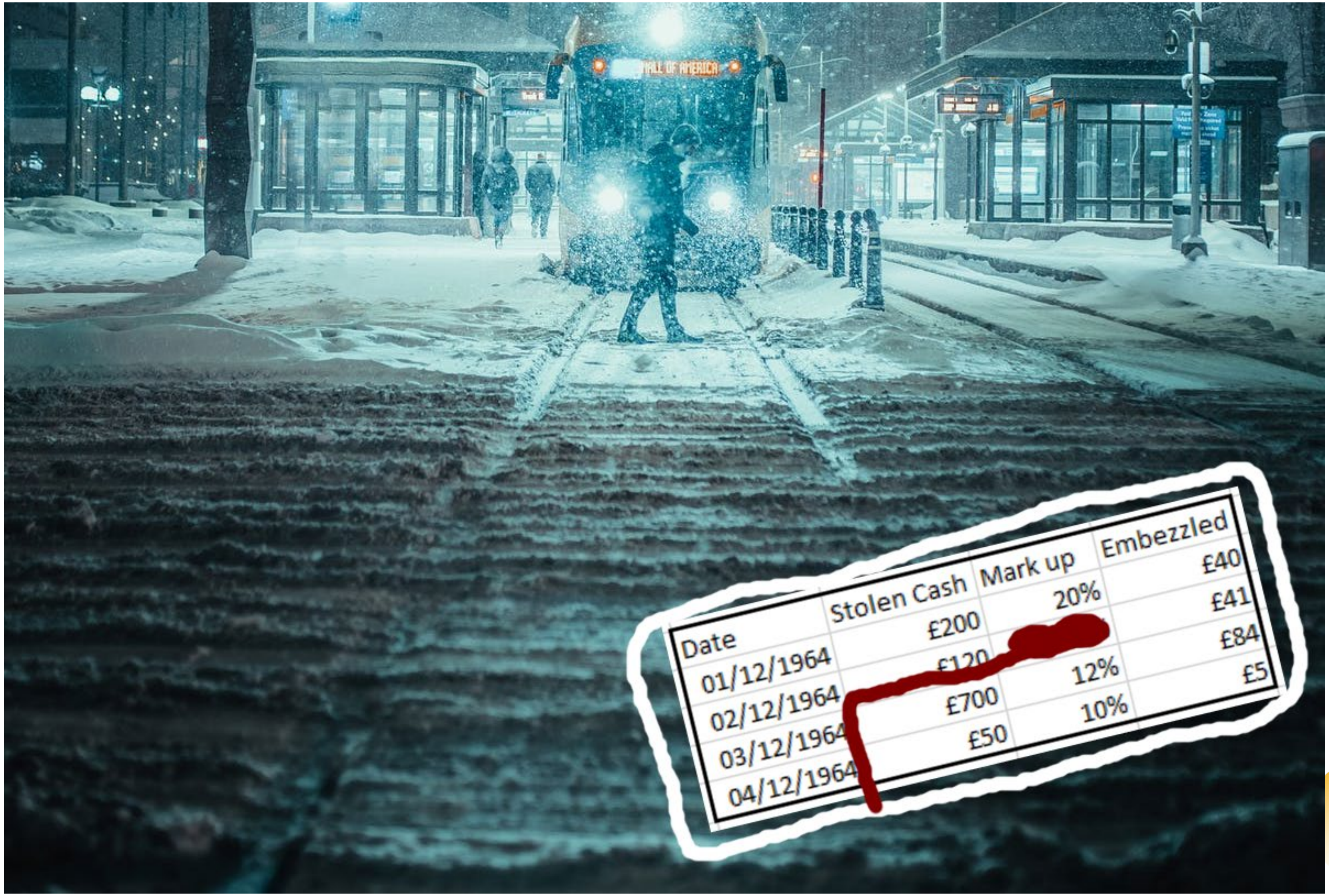
Git terminology

Terms to learn

- Repository – a place to store a project
- Cloning – making a local copy of a repository
- Pulling – update your local copy of a repository
- Pushing – sending your local copy to the repo
- Conflicting – when two users push conflicting versions
- Committing – captures changes with a friendly message
- Adding – add something to a commit
- Removing – remove something from a commit
- Status – the current state of our commit

Accountability on a crime scene

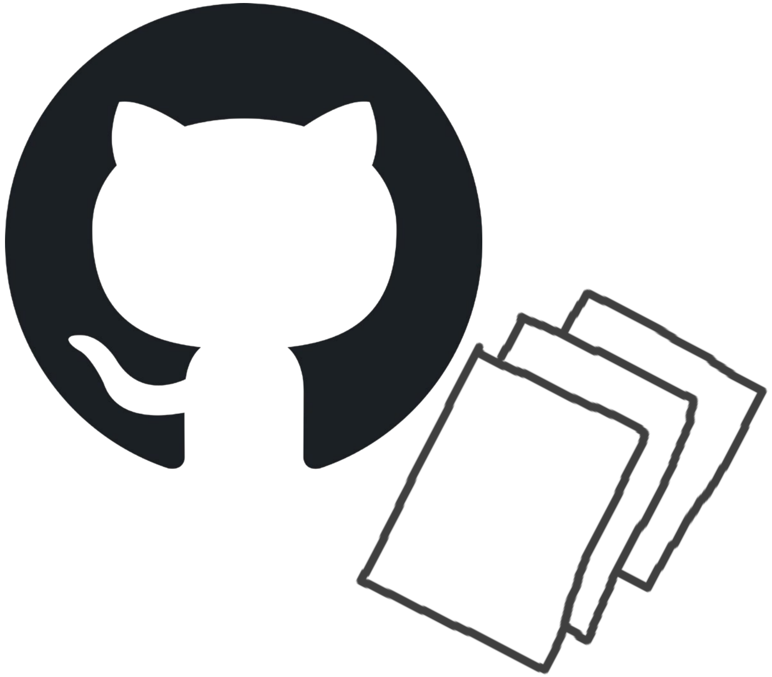
The Crime



Date	Stolen Cash	Mark up	Embezzled
01/12/1964	£200	20%	£40
02/12/1964	£120	[redacted]	£41
03/12/1964	£700	12%	£84
04/12/1964	£50	10%	£5

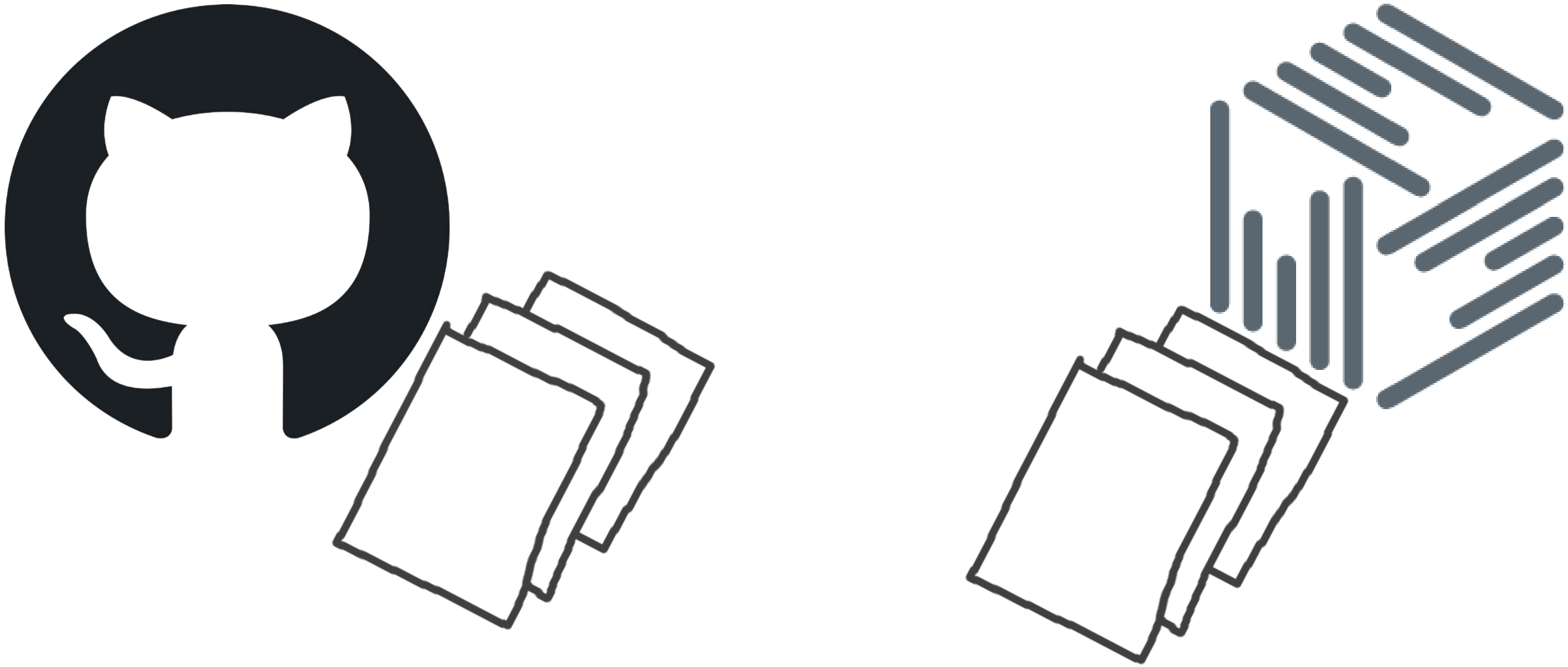
The Repository

- We could store our case notes in our office
- But the local police station – “The **Repository**” is a great place to store important files.



We **clone** the repository

- With a safe copy in the **Repository**, we can **clone** a full copy to take home.



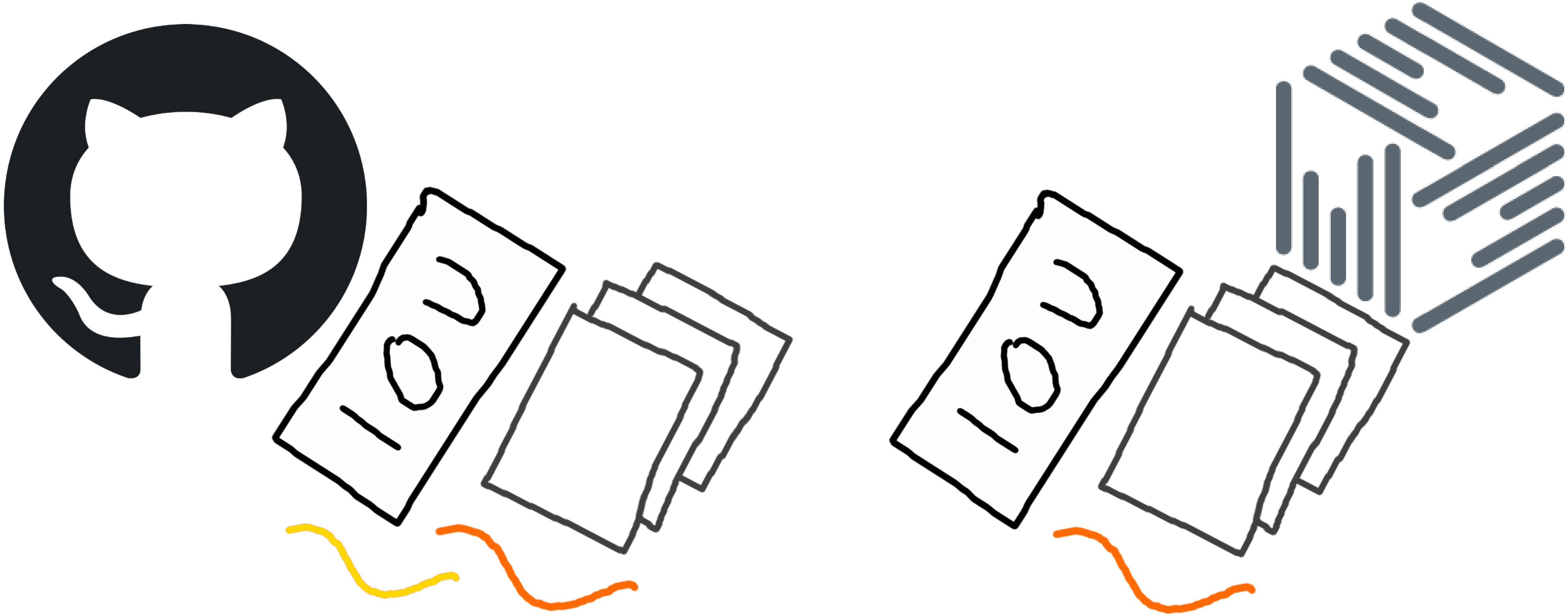
We return to the scene of the crime

We quickly **add** these to our notes

- We **add** these to our notes.
- Our **status** lists our changes:
  - + 1 IOU
  - + 1 Orange hair
- We write a **commit** in our notebook
  - **“Found evidence at the crime scene”**

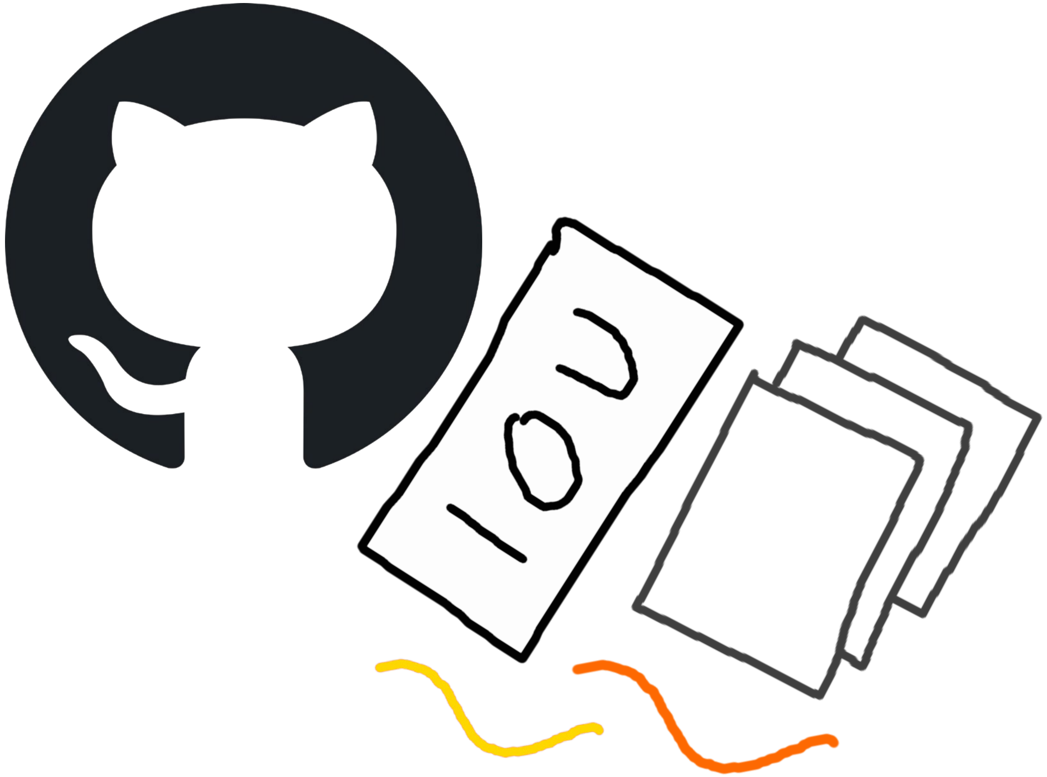
We **push** the repository

- Our notes are ahead of the **repository.**
- We **push** our notes, and **commit** message.



We **pull** the repository

- Our notes are ahead of the **repository.**
- The Repository is ahead of our notes
- We **pull** our notes, and **commit** message.

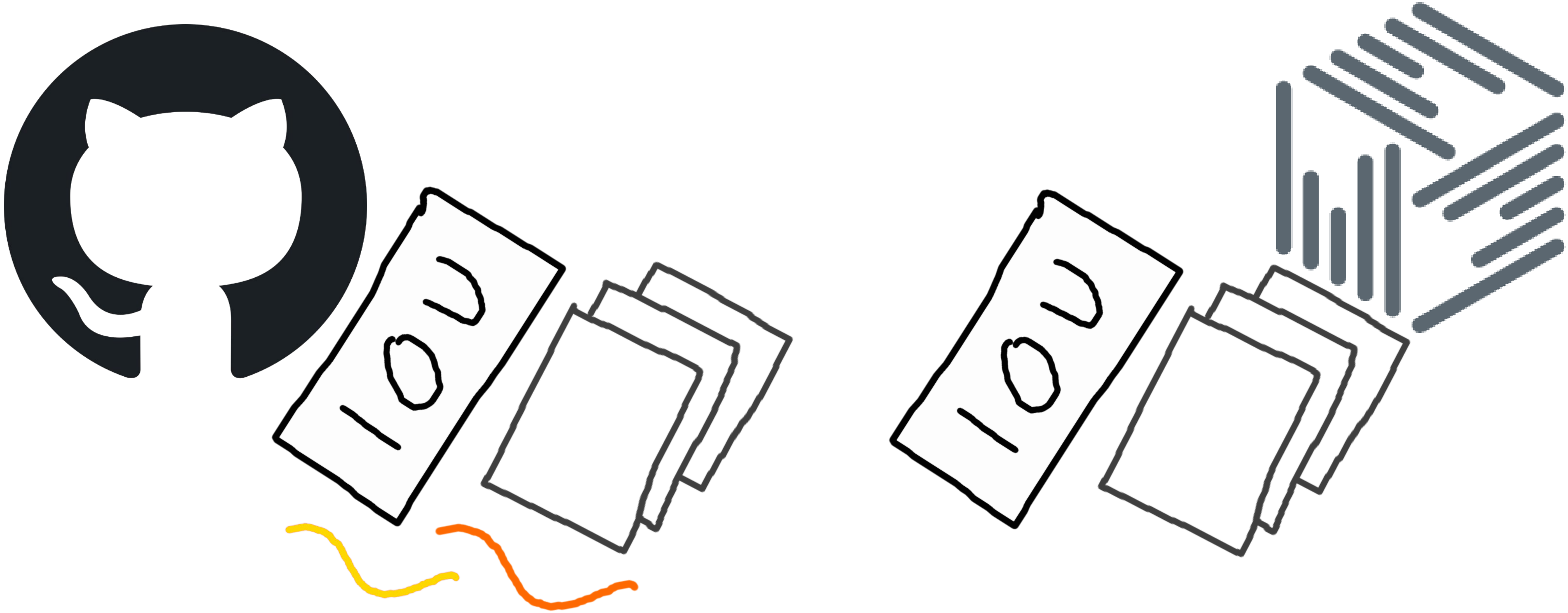


Investigate the alleyway

- We find a bunch more hair
- We find an orange cat…
- **Remove** our orange and blonde hairs
- Check the **status**
  - - 1 orange hair
  - - 1 yellow hair
- Write a **commit "Remove cat hair from the case"**

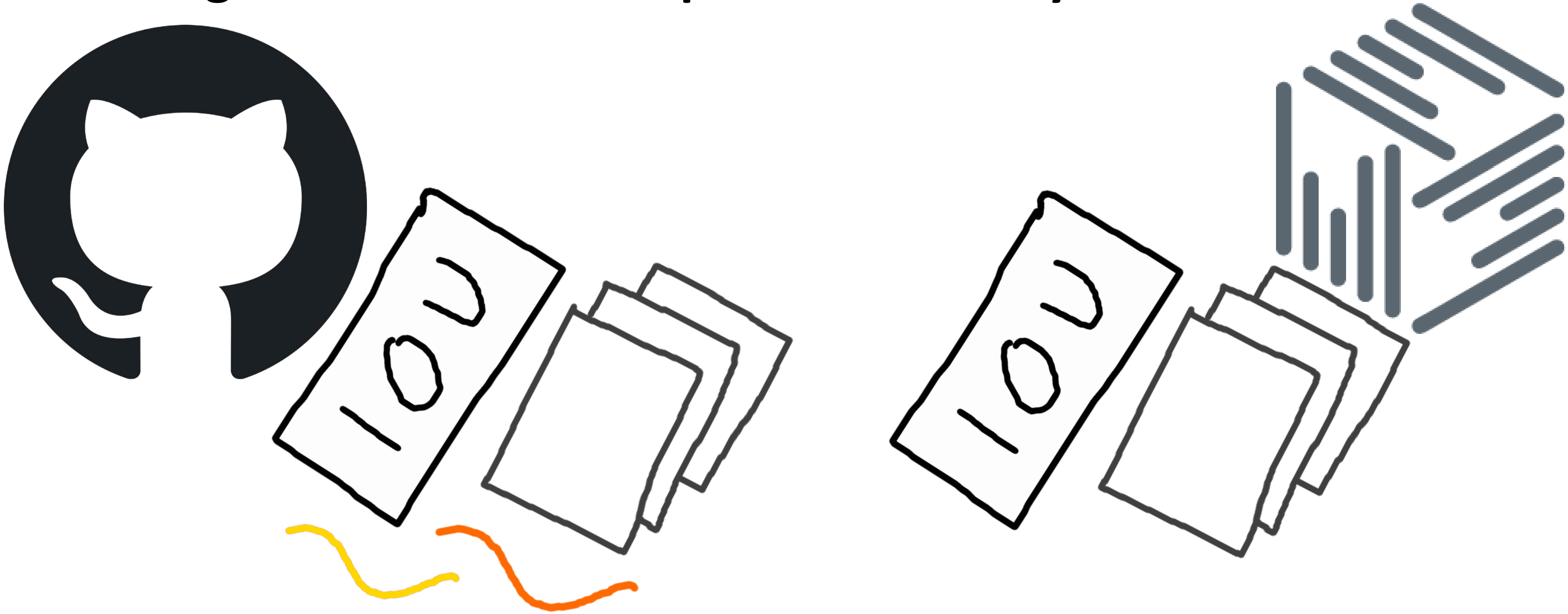
We **push** the repository

- Our notes are ahead of the **repository.**
- We **push** our notes, and **commit** message.



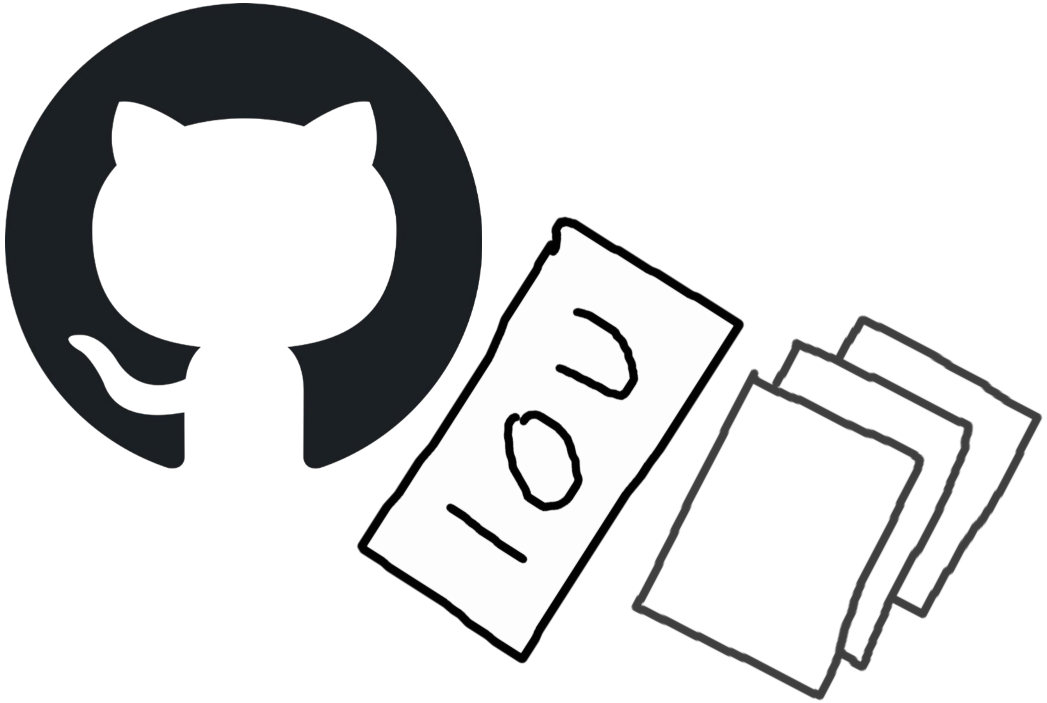
A Conflict

- When two detectives push contradictory changes
- We say **“Remove cat hair from the case”**
- **“orange and blonde hairs point to the mayor!”**



Case closed

- The second detective was using the cat hairs to frame the mayor!
- Luckily our work is so reproducible!



GitHub demo – writing an English paper, together.

Exercises

1. Create a new Repository
2. Make a new file, commit and push.
3. Make an edit to a file, commit and push.
4. Clone the Repository

Writing an English paper

Building an essay

- 1. Create a new repository
- Show README, add to readme
- 2. Write notes.txt, write a good commit message and save
- 3. Enhance notes.txt, write a good commit message and save
- Delete repo(REMOVE THESE TITLES)
- 4. Write essay.txt, flesh out point 1 and point 2, write good commit message
- 5. Write a conclusion, spell andy wrong
- 6. Write an introduction
- 7. Review and save first draft
- 8. proof read again
- 9. Send copy so far to review, if my friend or teacher knows version control they can do this within my system.
- Revert teachers notes
- Conflict

Did we fix what went wrong?

- Overwritten files
- Context hidden in filenames
- Deletion of work
- Single fail point
- Collaboration creates complexity
- No accountability we did any work

Git good

- Commit as often as you can, see this as saving but with context
- Write a README as if you’ve never seen your project before
- Visualize a tree

Next steps?

- Try using GitHub to store some of your files

Want more?

- An example exercise - gcapes.github.io/git-course/03-history/#exercise-bio-repository
- Interactive tutorial - learngitbranching.js.org
- Reproducibility and why it matters for you www.youtube.com/watch?v=RDCHTJEOV7g

Sources

- More technical - gcapes.github.io/git-course/slideshow/index.html
- Chambers 2014 - www.theguardian.com/science/head-quarters/2014/jun/10/physics-envy-do-hard-sciences-hold-the-solution-to-the-replication-crisis-in-psychology

Any questions?

UK Data Service

Thank you.

Joseph Allen

@JosephAllen1234

Joseph.Allen@Manchester.ac.uk