RESEARCH DATA MANAGEMENT

WITH DATALAD

Adina Wagner
@AdinaKrik

Lennart Wittkuhn
@lnnrtwttkhn

Psychoinformatics lab,
Institute of Neuroscience and Medicine (INM-7)
Research Center Jülich

Max Planck Research Group NeuroCode,
Max Planck Institute for Human Development, Berlin
IMPRS COMP2PSYCH

Slides: https://github.com/datalad-handbook/course/

WELCOME!

A few logistical things first:

WELCOME!

A few logistical things first:

- An approximate schedule for today is on our companion workshop website (link is in the public notes). We'll try to stick to it

WELCOME!

A few logistical things first:

- An approximate schedule for today is on our companion workshop website (link is in the public notes). We'll try to stick to it
- Let us introduce the workshop organizers...

WELCOME!

A few logistical things first:

- An approximate schedule for today is on our companion workshop website (link is in the public notes). We'll try to stick to it
- Let us introduce the workshop organizers...
- Let us introduce the virtual workshop venue...

WELCOME!

A few logistical things first:

- An approximate schedule for today is on our companion workshop website (link is in the public notes). We'll try to stick to it
- Let us introduce the workshop organizers...
- Let us introduce the virtual workshop venue...
    - Public and private chats

WELCOME!

A few logistical things first:

- An approximate schedule for today is on our companion workshop website (link is in the public notes). We'll try to stick to it
- Let us introduce the workshop organizers...
- Let us introduce the virtual workshop venue...
  - Public and private chats
  - Shared notes

WELCOME!

A few logistical things first:

- An approximate schedule for today is on our companion workshop website (link is in the public notes). We'll try to stick to it
- Let us introduce the workshop organizers...
- Let us introduce the virtual workshop venue...
  - Public and private chats
  - Shared notes
  - Break-out rooms

WELCOME!

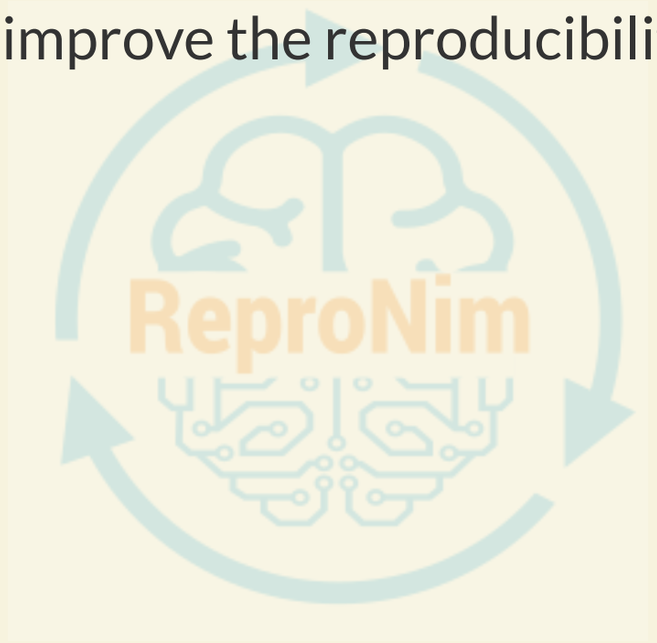
A few logistical things first:

- An approximate schedule for today is on our companion workshop website (link is in the public notes). We'll try to stick to it
- Let us introduce the workshop organizers...
- Let us introduce the virtual workshop venue...
  - Public and private chats
  - Shared notes
  - Break-out rooms
  - Drop out and re-join as you please, make use of your status setting

WHY ARE WE HERE? REPRONIM

WHY ARE WE HERE? REPRONIM

- ReproNim is an initiative to improve the reproducibility and efficiency in neuroimaging



WHY ARE WE HERE? REPRONIM

- ReproNim is an initiative to improve the reproducibility and efficiency in neuroimaging
- It's goal is "to improve the reproducibility of neuroimaging science and extend the value of our national investment in neuroimaging research, while making the process easier and more efficient for investigators."

WHY ARE WE HERE? REPRONIM

- ReproNim is an initiative to improve the reproducibility and efficiency in neuroimaging
- It's goal is "to improve the reproducibility of neuroimaging science and extend the value of our national investment in neuroimaging research, while making the process easier and more efficient for investigators."
- ReproNim develops free training materials, supports tool development, and offers training activities. ReproNim fellows teach their peers in independent courses, workshops, or Hackathons about tools or methods that increase the reproducibility of their research.

WHAT WILL WE DO TODAY?

WHAT WILL WE DO TODAY?

- The workshop centers around DataLad

WHAT WILL WE DO TODAY?

- The workshop centers around DataLad
- We aim to do more than a standard introduction by providing in-depth explanations, hands-on exercises, and discussions throughout the day

WHAT WILL WE DO TODAY?

- The workshop centers around DataLad
- We aim to do more than a standard introduction by providing in-depth explanations, hands-on exercises, and discussions throughout the day
- (this will be much harder in this virtual setting - please bear with us)

LET'S DO THE SPLITS

QUESTIONS/INTERACTION THROUGHOUT THE WORKSHOP

- If you have a question during a lecture, please first type your questions in the chat. There are no stupid questions :)
- It would be great to have lively discussions - unless its interrupting a speaker, please feel encouraged to unmute/turn on your video to interact with us.
- We're happy to discuss specific use cases at the end. Please make a note about them in the "Shared notes" of BigBlueButton
- We are recording the lectures and will make them available online

QUESTIONS/INTERACTION AFTER THE WORKSHOP

If you have a question after the workshop, you can reach out for help:

Reach out to to the DataLad team via

- Matrix (free, decentralized communication app, no app needed)
- or the development repository on GitHub

Reach out to the user community with

- A question on neurostars.org with a `datalad` tag

Find more user tutorials and workshop recordings

- On DataLad's YouTube channel
- In the DataLad Handbook

RESOURCES AND FURTHER READING

Comprehensive user documentation in the DataLad Handbook (handbook.datalad.org)

- High-level function/command overviews, Installation, Configuration, Cheatsheet

---

- Narrative-based code-along course
- Independent on background/skill level, suitable for data management novices

---

- Step-by-step solutions to common data management problems, like how to make a reproducible paper

LIVE POLLING SYSTEM

http://etc.ch/5xo7

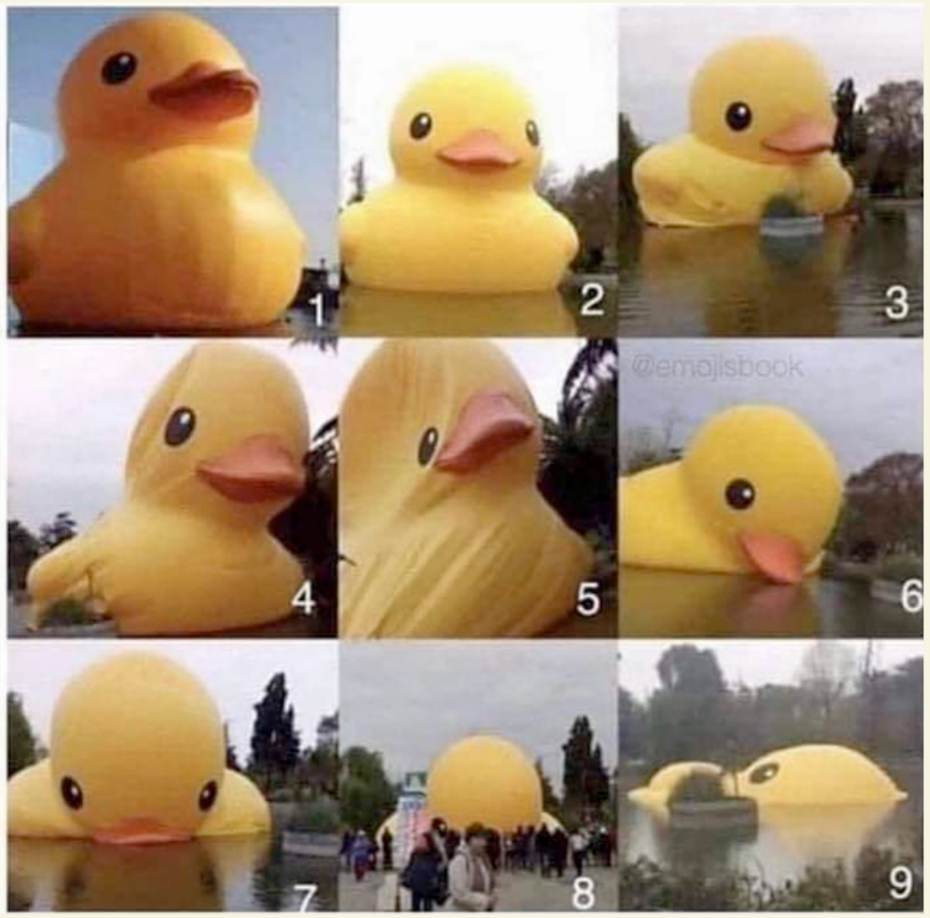
Direct Poll

WHAT'S YOUR MOOD TODAY?

http://etc.ch

Direct Poll

WHAT'S YOUR MOOD TODAY?



http://etc.ch

Direct Poll

HELP US LEARN MORE ABOUT OUR AUDIENCE

http://etc.ch/5xo7

Direct Poll

LIVE CODING + HANDS-ON

- Live-demonstration of DataLad examples and workflows
- Code along with copy-paste code snippets and hands-on exercises on the workshop website
- Requirements:
  - Most recent DataLad version for your OS (installation instructions at handbook.datalad.org)
  - For containerized analyses: DataLad extension datalad-containers (available via pip) + Singularity or Docker

MOTIVATION

REAL WORLD EXAMPLES FOR RESEARCH DATA MANAGEMENT GONE WRONG ...

A million-row limit on Microsoft's Excel spreadsheet software may have led to Public Health England misplacing nearly 16,000 Covid test results, it is understood.

The data error, which led to 15,841 positive tests being left off the official daily figures, means than 50,000 potentially infectious people may have been missed by contact tracers and not told to self-isolate.

In this case, the Guardian understands, one lab had sent its daily test report to PHE in the form of a CSV file - the simplest possible database format, just a list of values separated by commas. That report was then loaded into Microsoft Excel, and the new tests at the bottom were added to the main database.

But while CSV files can be any size, Microsoft Excel files can only be 1,048,576 rows long - or, in older versions which PHE may have still been using, a mere 65,536. When a CSV file longer than that is opened, the bottom rows get cut off and are no longer displayed. That means that, once the lab had performed more than a million tests, it was only a matter of time before its reports failed to be read by PHE.

https://www.theguardian.com/politics/2020/oct/05/how-excel-may-have-caused-loss-of-16000-covid-tests-in-england

REAL WORLD EXAMPLES FOR RESEARCH DATA MANAGAMENT GONE WRONG ...

Reinhart and Rogoff's work showed average real economic growth slows (a 0.1% decline) when a country's debt rises to more than 90% of gross domestic product (GDP) – and this 90% figure was employed repeatedly in political arguments over high-profile austerity measures.

The most serious was that, in their Excel spreadsheet, Reinhart and Rogoff had not selected the entire row when averaging growth figures: they omitted data from Australia, Austria, Belgium, Canada and Denmark.

So the key conclusion of a seminal paper, which has been widely quoted in political debates in North America, Europe Australia and elsewhere, was invalid.

While many different types of errors were involved in these calamities, the fact that the errors in the Reinhart-Rogoff paper were not identified earlier can be ascribed by the pervasive failure of scientific and other researchers to make all data and computer code publicly available at an early stage – preferably when the research paper documenting the study is submitted for review.

https://theconversation.com/the-reinhart-rogoff-error-or-how-not-to-excel-at-economics-13646

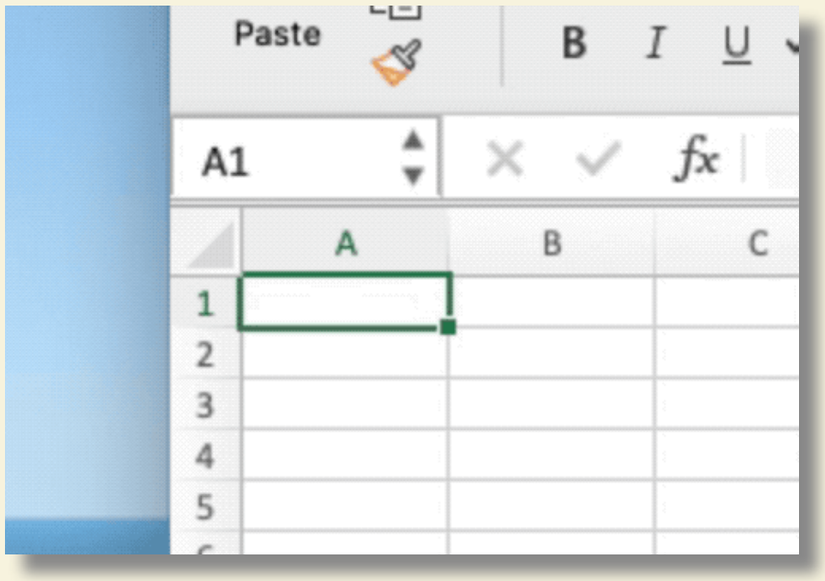
REAL WORLD EXAMPLES FOR RESEARCH DATA MANAGAMENT GONE WRONG ...

MICROSOFT | REPORT | SCIENCE

Scientists rename human genes to stop Microsoft Excel from misreading them as dates

Sometimes it's easier to rewrite genetics than update Excel

By James Vincent | Aug 6, 2020, 8:44am EDT



Help has arrived, though, in the form of the scientific body in charge of standardizing the names of genes, the HUGO Gene Nomenclature Committee, or HGNC. This week, the HGNC published new guidelines for gene naming, including for "symbols that affect data handling and retrieval." From now on, they say, human genes and the proteins they expressed will be named with one eye on Excel's auto-formatting. That means the symbol MARCH1 has now become MARCHF1, while SEPT1 has become SEPTIN1, and so on. A record of old symbols and names will be stored by HGNC to avoid confusion in the future.

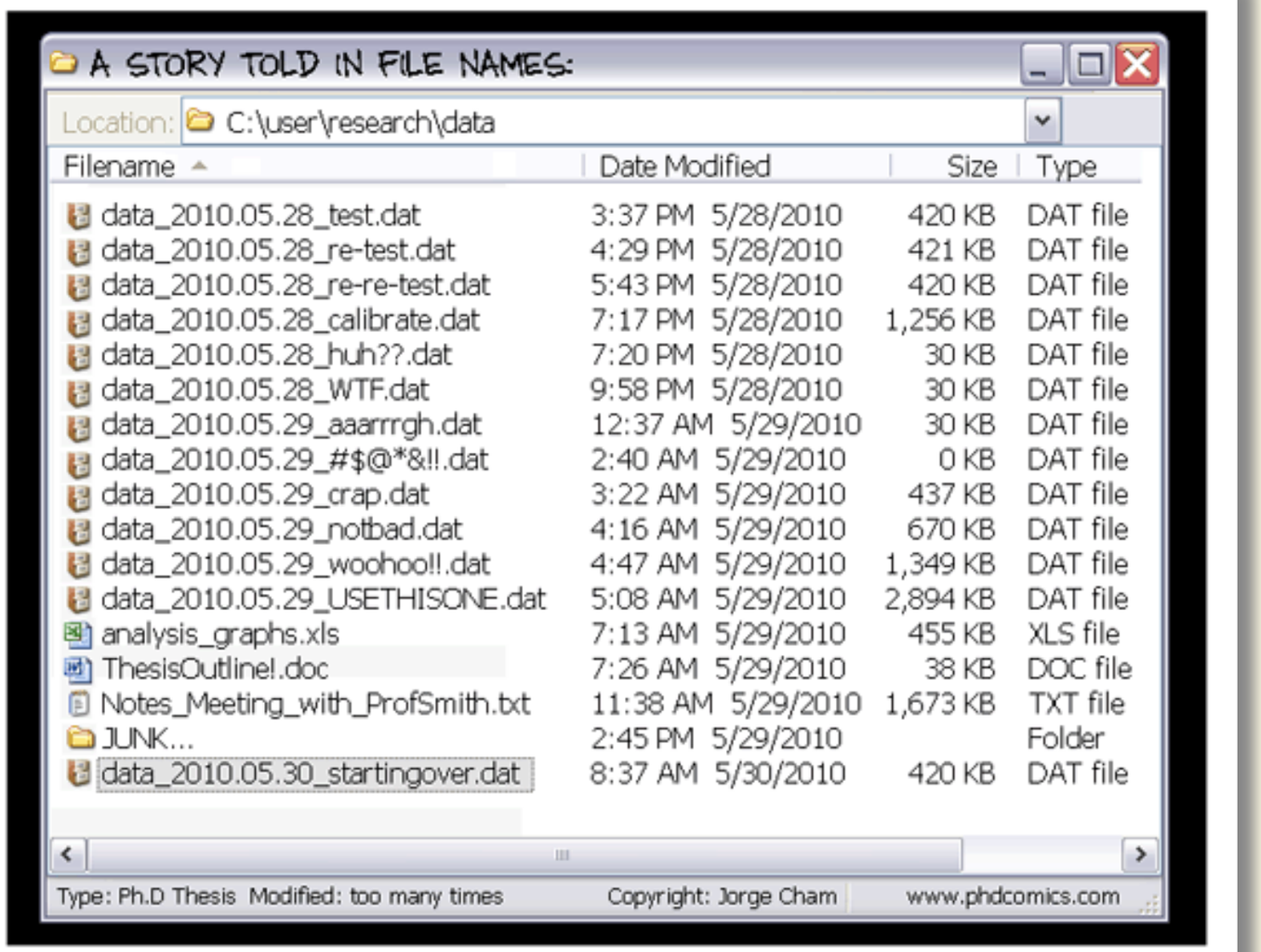
https://www.theverge.com/2020/8/6/21355674/human-genes-rename-microsoft-excel-misreading-dates

REAL WORLD EXAMPLE CLOSER TO HOME:
"REPLICATION CRISIS" IN PSYCHOLOGY / NEUROSCIENCE

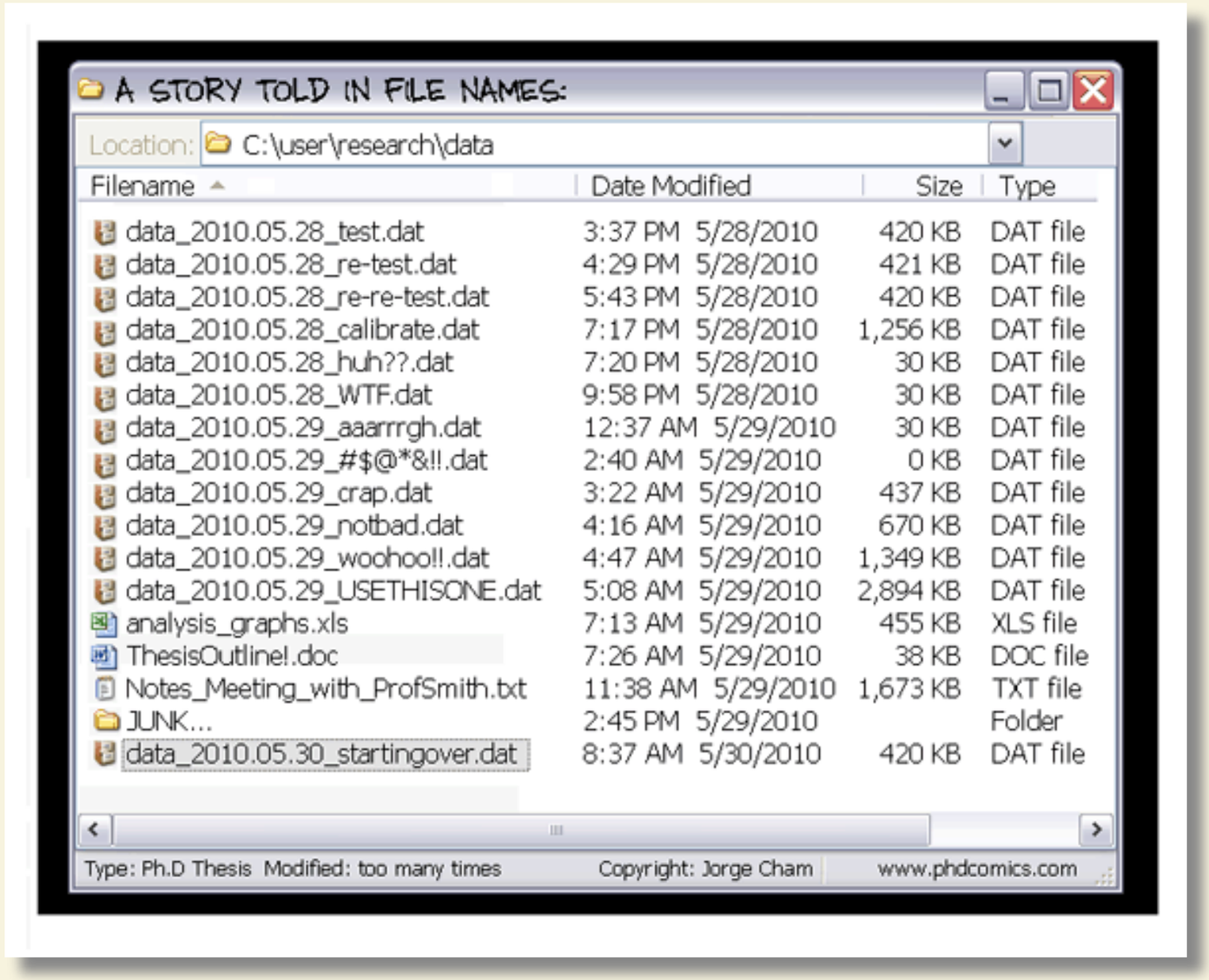
Data from many fields suggests reproducibility is lower than is desirable[8–14]; one analysis estimates that 85% of biomedical research efforts are wasted[14], while 90% of respondents to a recent survey in *Nature* agreed that there is a 'reproducibility crisis'[15]. Whether 'crisis' is the appropriate term to describe the current state or trajectory of science is debatable, but accumulated evidence indicates that there is substantial room for improvement with regard to research practices to maximize the efficiency of the research community's use of the public's financial investment in research.

taken from "A manifesto for reproducible science" by Munafò et al., 2017, *Nature Human Behavior*

DATA CHANGE!

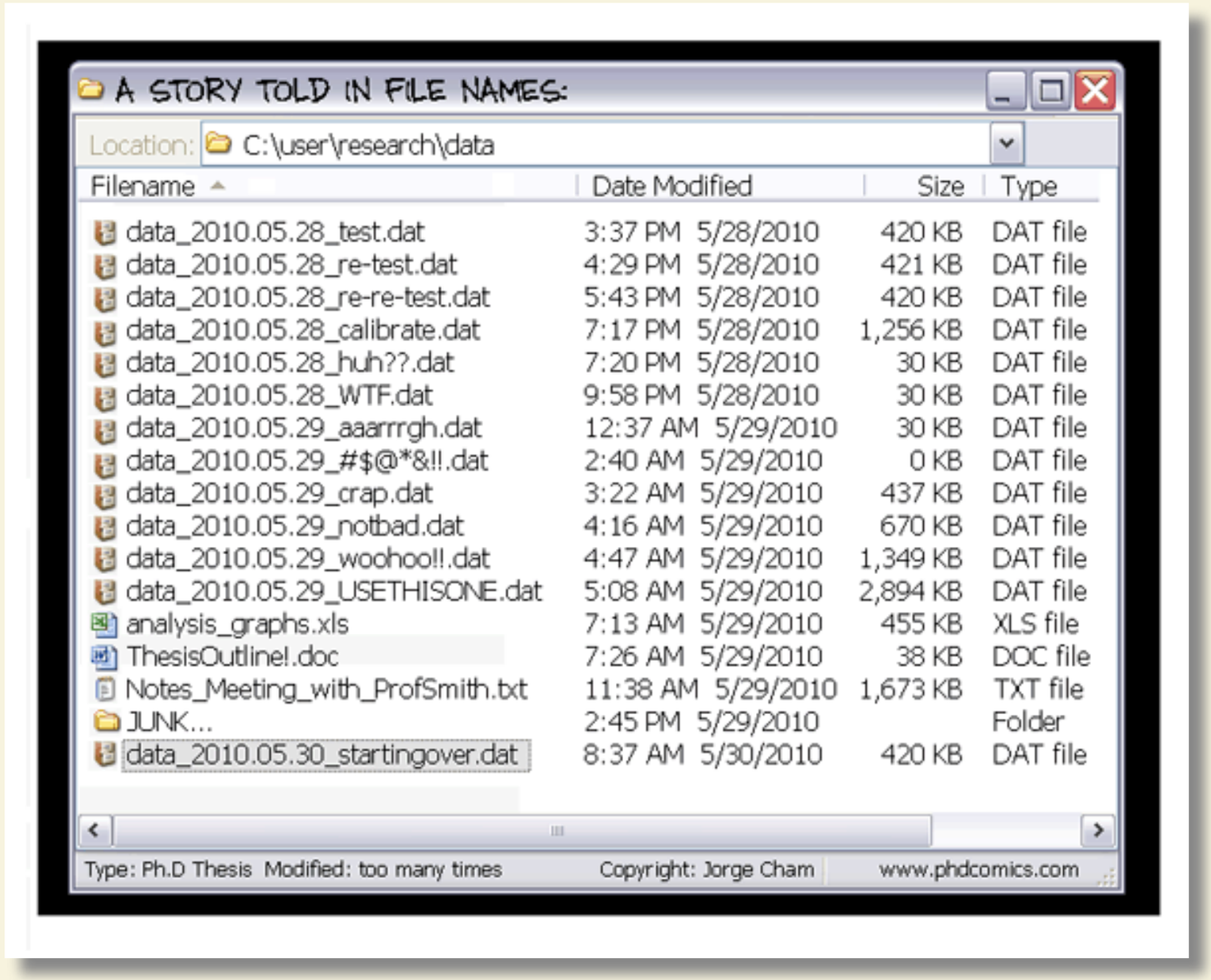

A STORY TOLD IN FILE NAMES:
Location: C:\user\research\data
Filename
Date Modified
Size
Type
data_2010.05.28_test.dat 3:37 PM 5/28/2010 420 KB DAT file
data_2010.05.28_re-test.dat 4:29 PM 5/28/2010 421 KB DAT file
data_2010.05.28_re-re-test.dat 5:43 PM 5/28/2010 420 KB DAT file
data_2010.05.28_calibrate.dat 7:17 PM 5/28/2010 1,256 KB DAT file
data_2010.05.28_huh??.dat 7:20 PM 5/28/2010 30 KB DAT file
data_2010.05.28_WTF.dat 9:58 PM 5/28/2010 30 KB DAT file
data_2010.05.29_aaarrrgh.dat 12:37 AM 5/29/2010 30 KB DAT file
data_2010.05.29_#$@*&!!.dat 2:40 AM 5/29/2010 0 KB DAT file
data_2010.05.29_crap.dat 3:22 AM 5/29/2010 437 KB DAT file
data_2010.05.29_notbad.dat 4:16 AM 5/29/2010 670 KB DAT file
data_2010.05.29_woohoo!!.dat 4:47 AM 5/29/2010 1,349 KB DAT file
data_2010.05.29_USETHISONE.dat 5:08 AM 5/29/2010 2,894 KB DAT file
analysis_graphs.xls 7:13 AM 5/29/2010 455 KB XLS file
ThesisOutline!.doc 7:26 AM 5/29/2010 38 KB DOC file
Notes_Meeting_with_ProfSmith.txt 11:38 AM 5/29/2010 1,673 KB TXT file
JUNK... 2:45 PM 5/29/2010 Folder
data_2010.05.30_startingover.dat 8:37 AM 5/30/2010 420 KB DAT file
Type: Ph.D Thesis Modified: too many times
Copyright: Jorge Cham
www.phdcomics.com

DATA CHANGE!



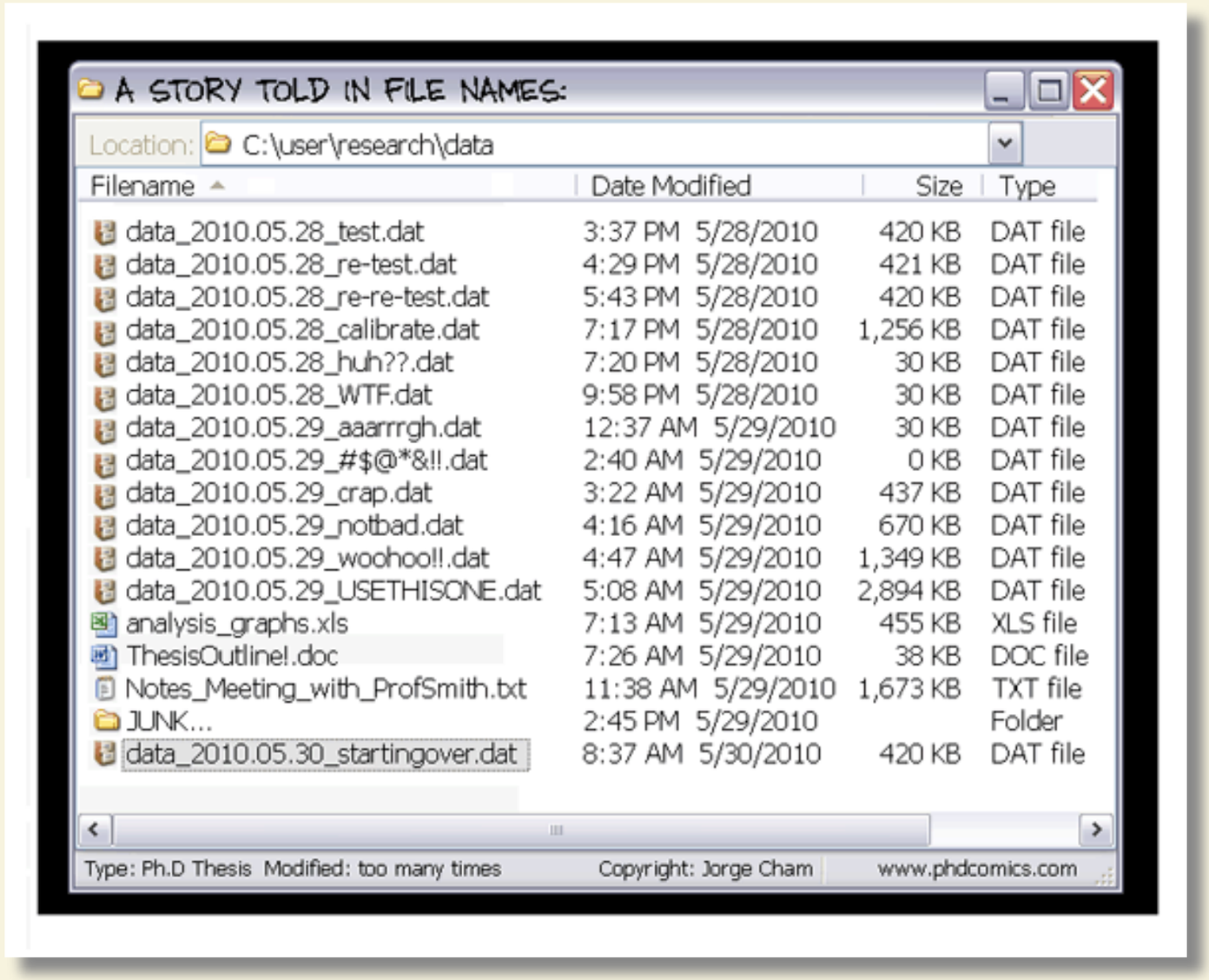
- New data are added and old data removed

DATA CHANGE!



- New data are added and old data removed
- Erros are detected, fixed and introduced again

DATA CHANGE!

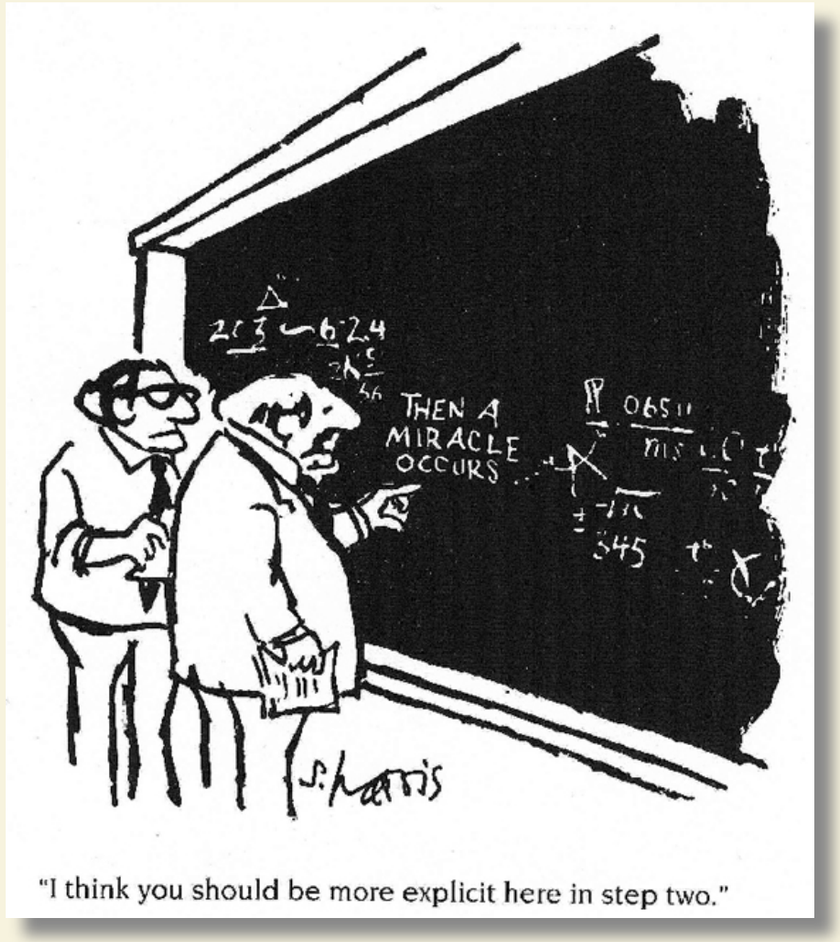


- New data are added and old data removed
- Erros are detected, fixed and introduced again
- Separate data versions are created or merged

METHODS DOCUMENTATION AND PROVENANCE

Analytic flexibility leads to sizeable variations in results

(see e.g., Carp. 2012 and Botvinik-Nezer, 2020 for examples from neuroimaging)

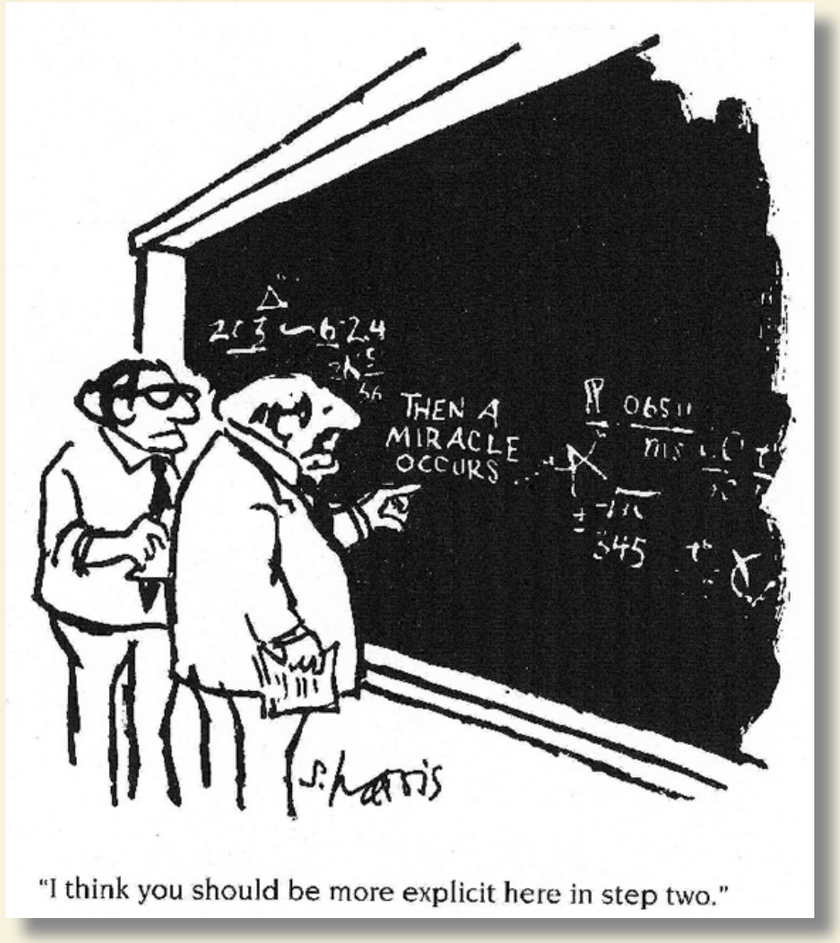


"I think you should be more explicit here in step two."

METHODS DOCUMENTATION AND PROVENANCE

Analytic flexibility leads to sizeable variations in results

(see e.g., Carp. 2012 and Botvinik-Nezer, 2020 for examples from neuroimaging)



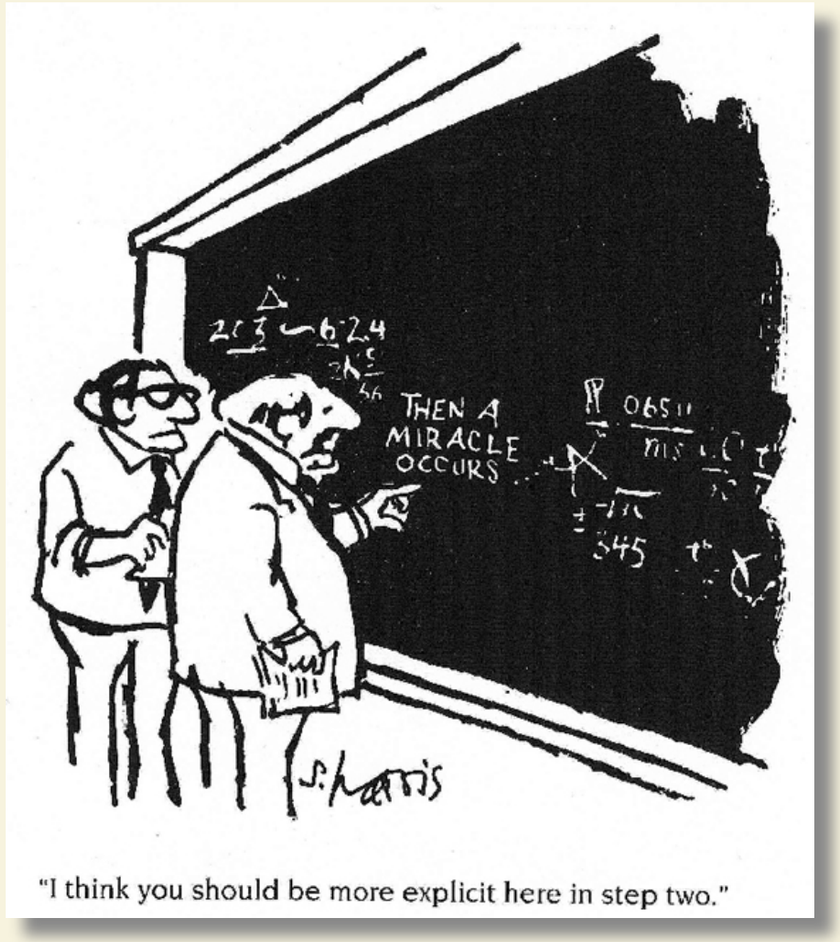
"I think you should be more explicit here in step two."

- provide information on how data came into existence

METHODS DOCUMENTATION AND PROVENANCE

Analytic flexibility leads to sizeable variations in results

(see e.g., Carp. 2012 and Botvinik-Nezer, 2020 for examples from neuroimaging)



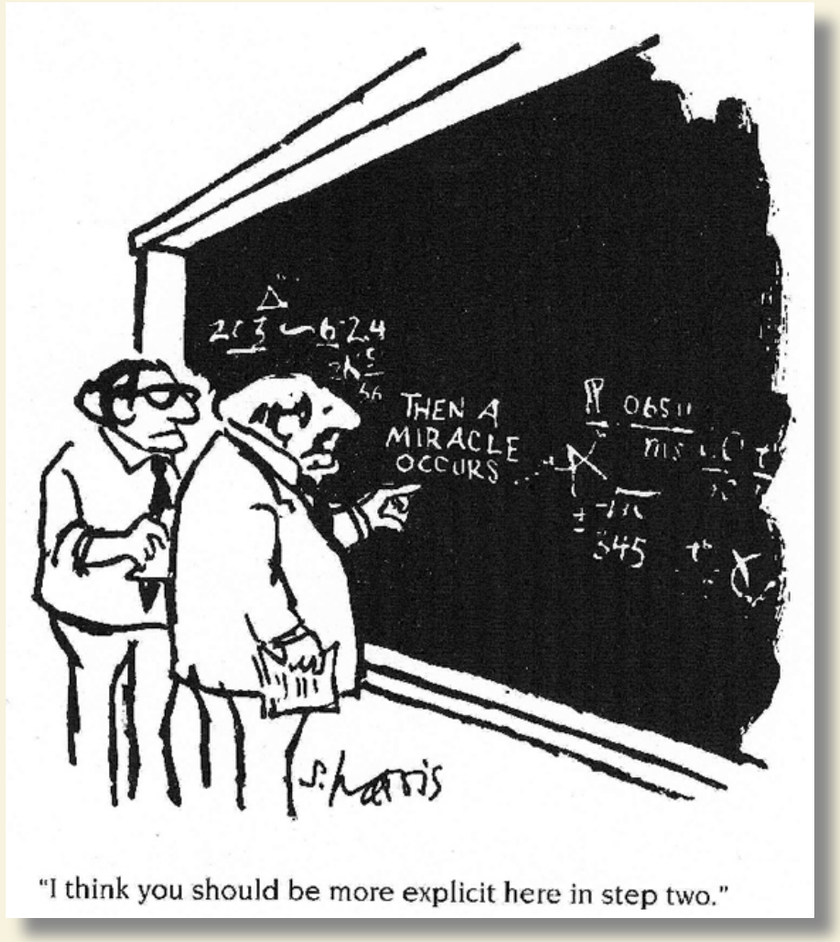
"I think you should be more explicit here in step two."

- provide information on how data came into existence
- change data through documented code, not manually

METHODS DOCUMENTATION AND PROVENANCE

Analytic flexibility leads to sizeable variations in results

(see e.g., Carp. 2012 and Botvinik-Nezer, 2020 for examples from neuroimaging)



"I think you should be more explicit here in step two."

- provide information on how data came into existence
- change data through documented code, not manually
- relate changes in data to changes in code

WHAT'S HOLDING YOU BACK?

Hand on heart: Have you heard / said / thought these statements* before?

1. "It's only the paper / results that matters"
2. "I'd rather do real science than tidy up my data"
3. "Mind your own business! I document my data the way I want!"
4. "Reproducibility sounds alright, but my code and data are spread over so many hard drives and directories that it would just be too much work to collect them all in one place"
5. "We can always sort out the code and data after submission"
6. "My field is very competitive and I can't risk wasting time"

http://etc.ch/5xo7

Direct Poll

* cf. Markowetz, 2015

BUT WHAT'S IN IT FOR ME? "SELFISH" REASONS FOR REPRODUCIBILITY

"[...] science is all about more publications, more impact factor, more money and more career. More, more, more ...
So how does working reproducibly help me achieve more as a scientist." - Markowetz, 2015

see e.g., Markowetz, 2015, *Genome Biology*; Poldrack, 2019, *Neuron*

BUT WHAT'S IN IT FOR ME? "SELFISH" REASONS FOR REPRODUCIBILITY

"[...] science is all about more publications, more impact factor, more money and more career. More, more, more ...
So how does working reproducibly help me achieve more as a scientist." - Markowetz, 2015

- You want to avoid the disaster of publishing "a miracle"
- You will be faster (in the long run)
    - Finding and fixing errors will be faster
    - Progress on new projects will happen faster
- Researchers (reviewers!) will have more trust in your findings
- Data sharing can foster collaboration (with your past self, inside and outside your institution) and lead to new projects and publications
- You acquire (technical) skills that will likely become increasingly important for your career, either in academia or industry

see e.g., Markowetz, 2015, *Genome Biology*; Poldrack, 2019, *Neuron*

BUT WHAT'S IN IT FOR ME? "SELFISH" REASONS FOR REPRODUCIBILITY

"[...] science is all about more publications, more impact factor, more money and more career. More, more, more ...
So how does working reproducibly help me achieve more as a scientist." - Markowetz, 2015

- You want to avoid the disaster of publishing "a miracle"
- You will be faster (in the long run)
    - Finding and fixing errors will be faster
    - Progress on new projects will happen faster
- Researchers (reviewers!) will have more trust in your findings
- Data sharing can foster collaboration (with your past self, inside and outside your institution) and lead to new projects and publications
- You acquire (technical) skills that will likely become increasingly important for your career, either in academia or industry

It's just useful for your everyday work and makes your life easier!
(see next slides ...)

see e.g., Markowetz, 2015, *Genome Biology*; Poldrack, 2019, *Neuron*

COMMON PROBLEMS IN SCIENCE

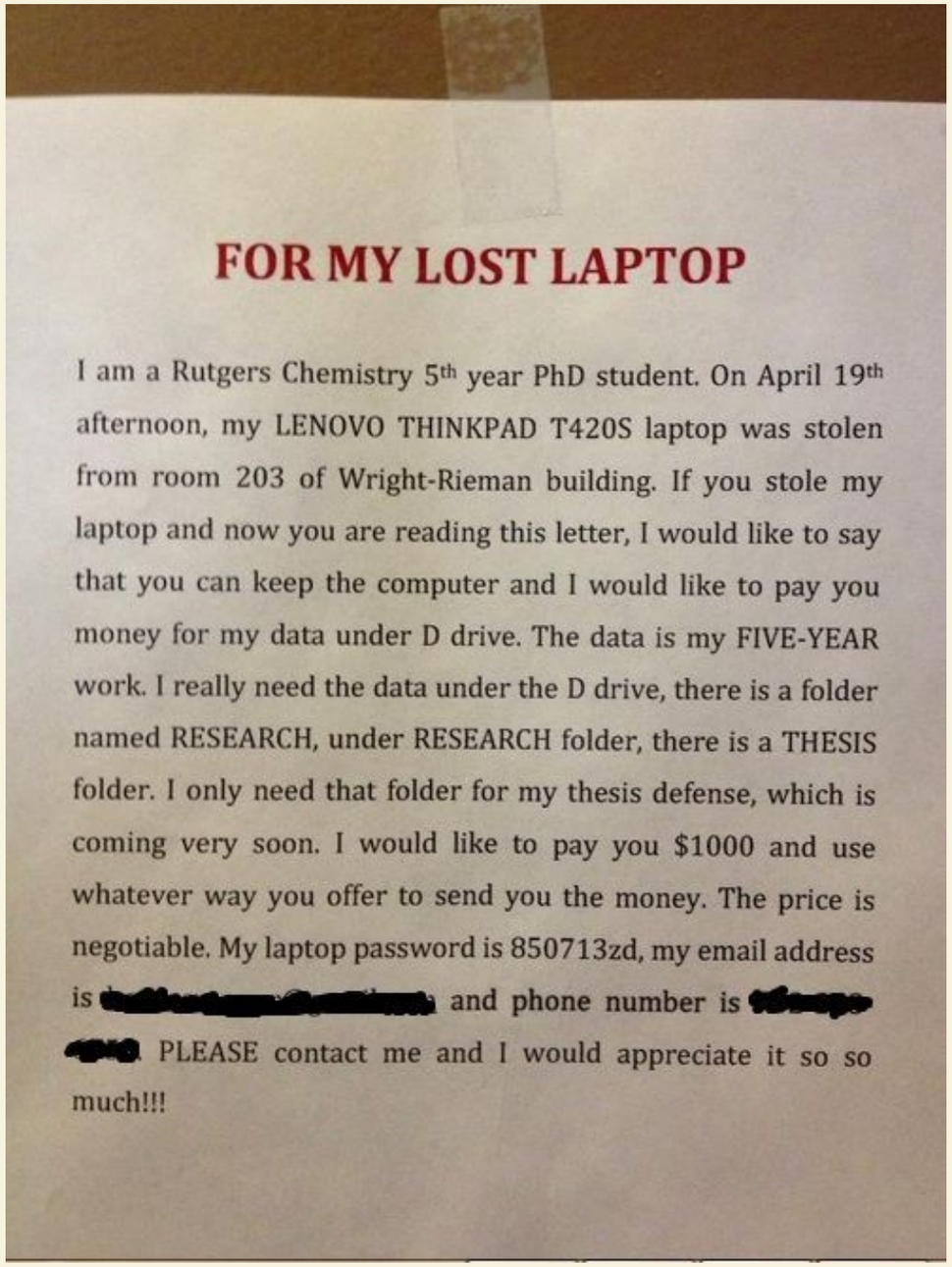
COMMON PROBLEMS IN SCIENCE

You write a paper about an algorithm, stay up late to generate good-looking figures, but you have to tweak parameters and display options to make it work AND look good. The next morning, you have no idea which parameters produced which figures, and which of the figures fits to what you report in the paper.

COMMON PROBLEMS IN SCIENCE

COMMON PROBLEMS IN SCIENCE

Your research project produces phenomenal results, but your laptop, the only place that stores the source code for the results, is stolen/breaks



FOR MY LOST LAPTOP

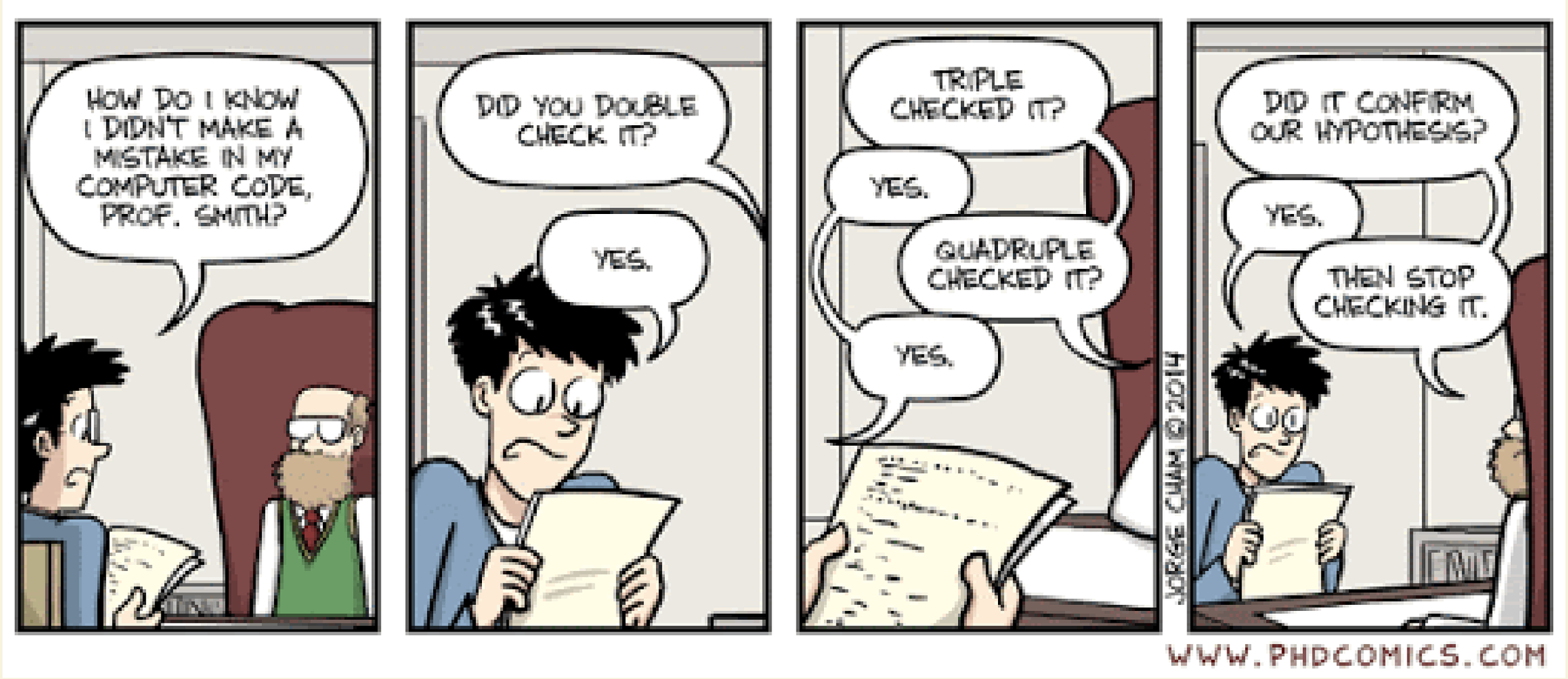
I am a Rutgers Chemistry 5th year PhD student. On April 19th afternoon, my LENOVO THINKPAD T420S laptop was stolen from room 203 of Wright-Rieman building. If you stole my laptop and now you are reading this letter, I would like to say that you can keep the computer and I would like to pay you money for my data under D drive. The data is my FIVE-YEAR work. I really need the data under the D drive, there is a folder named RESEARCH, under RESEARCH folder, there is a THESIS folder. I only need that folder for my thesis defense, which is coming very soon. I would like to pay you $1000 and use whatever way you offer to send you the money. The price is negotiable. My laptop password is 850713zd, my email address is [redacted] and phone number is [redacted] PLEASE contact me and I would appreciate it so so much!!!

COMMON PROBLEMS IN SCIENCE

Image credit: http://phdcomics.com/comics.php?f=1693

COMMON PROBLEMS IN SCIENCE

A graduate student approaches their supervisor, complaining that the supervisors research idea does not work. After weeks of discussion, it becomes apparent that oral communication doesn't suffice - the student can't sufficiently explain the environment (data, algorithms, ...) they constructed, and if the supervisor can't enter and use the students project there's no way to find a fix.



COMMON PROBLEMS IN SCIENCE

Image credit: http://phdcomics.com/comics.php?f=1693

COMMON PROBLEMS IN SCIENCE

A Post-doc wrote a script during the PhD that applied a specific method to a dataset. Now, with new data and a new project, they try to reuse the script, but forgot how it worked.

COMMON PROBLEMS IN SCIENCE

Image credit: http://phdcomics.com/comics.php?f=1693

COMMON PROBLEMS IN SCIENCE

You try to recreate results from another lab's published paper. You base your re-implementation on everything reported in their paper, but the results you obtain look nowhere like the original.

~~COMMON~~ OLD PROBLEMS IN SCIENCE

~~COMMON~~ OLD PROBLEMS IN SCIENCE

All these problems were paraphrased from Buckheit & Donoho, **1995**

~~COMMON~~ OLD PROBLEMS IN SCIENCE

All these problems were paraphrased from Buckheit & Donoho, **1995**
Why don't we make our live easier?

~~COMMON~~ OLD PROBLEMS IN SCIENCE

All these problems were paraphrased from Buckheit & Donoho, **1995**
Why don't we make our live easier?
Both for you and your future self, as well as for science as a whole?

~~COMMON~~ OLD PROBLEMS IN SCIENCE

All these problems were paraphrased from Buckheit & Donoho, **1995**
Why don't we make our live easier?
Both for you and your future self, as well as for science as a whole?
The tools exist, and are getting easier and easier to use.

~~COMMON~~ OLD PROBLEMS IN SCIENCE

All these problems were paraphrased from Buckheit & Donoho, **1995**
Why don't we make our live easier?
Both for you and your future self, as well as for science as a whole?
The tools exist, and are getting easier and easier to use.
Sometimes, you only need to know that something exists and ...

~~COMMON~~ OLD PROBLEMS IN SCIENCE

All these problems were paraphrased from Buckheit & Donoho, **1995**
Why don't we make our live easier?
Both for you and your future self, as well as for science as a whole?
The tools exist, and are getting easier and easier to use.
Sometimes, you only need to know that something exists and ...

👋 just 👋 get 👋 started! 👋

~~COMMON~~ OLD PROBLEMS IN SCIENCE

All these problems were paraphrased from Buckheit & Donoho, **1995**
Why don't we make our live easier?
Both for you and your future self, as well as for science as a whole?
The tools exist, and are getting easier and easier to use.
Sometimes, you only need to know that something exists and ...

👋 just 👋 get 👋 started! 👋

... but also don't be too hard on yourself! 🤗

CONCEPTS

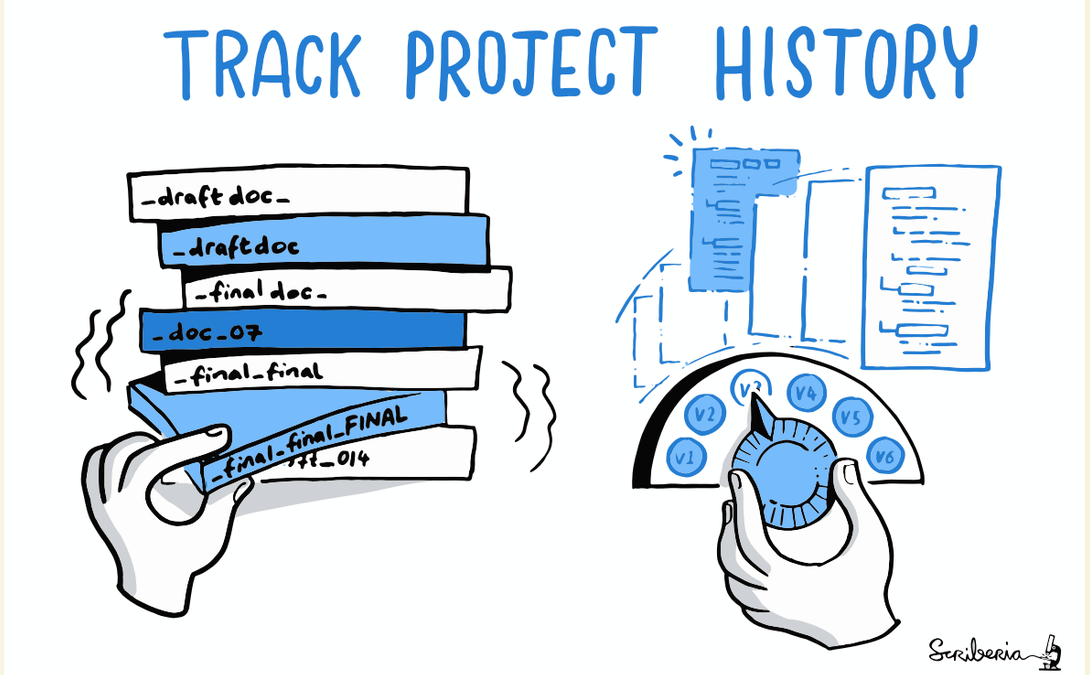
DEFINING REPLICABILITY

	Same data	New data
Same methods	Reproducibility	Replication
New methods	Robustness	Generalization

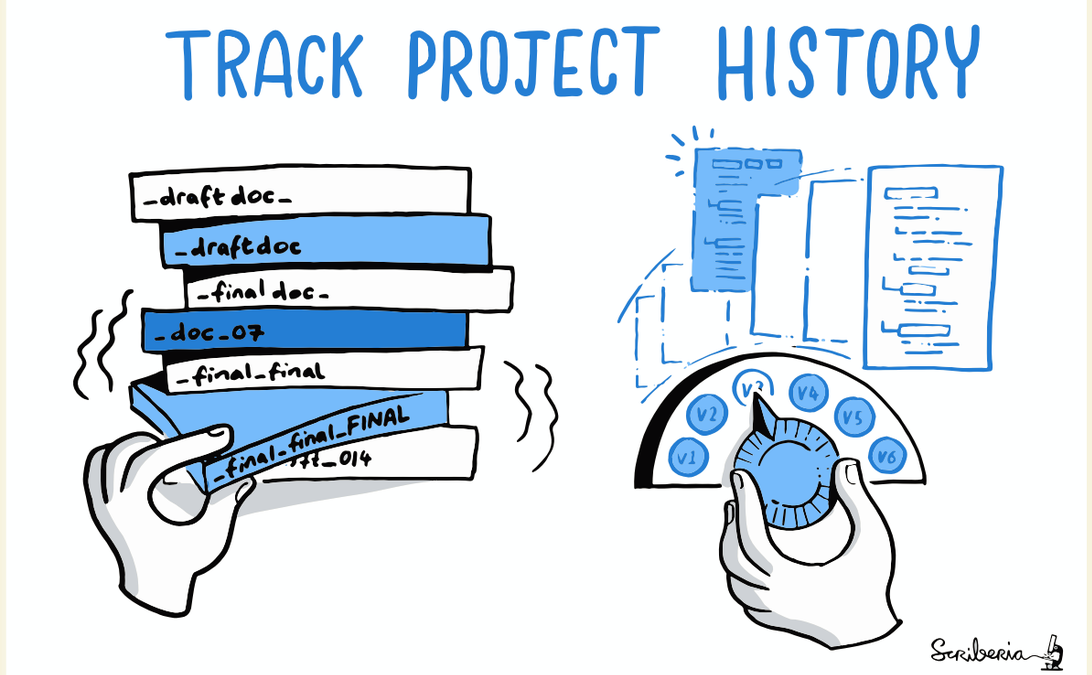
see e.g., Freese & Peterson, 2017

"Authors provide all the necessary data and the computer codes to run the analysis again, re-creating the results." - Claerbout & Karrenbach, **1992**

WHAT IS VERSION CONTROL?

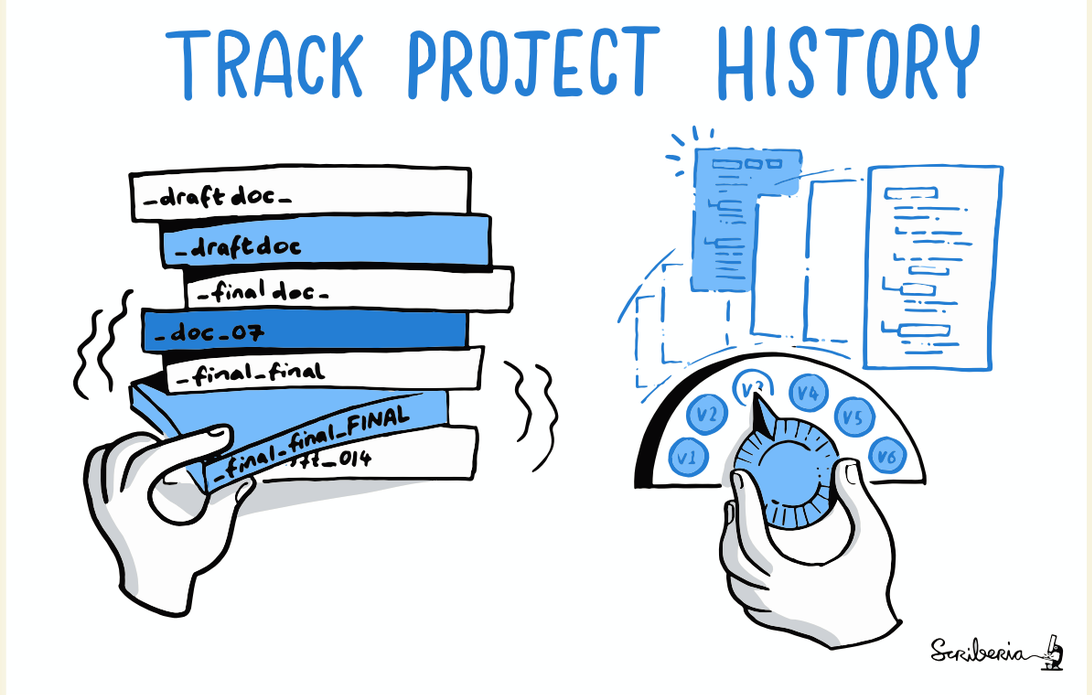


WHAT IS VERSION CONTROL?



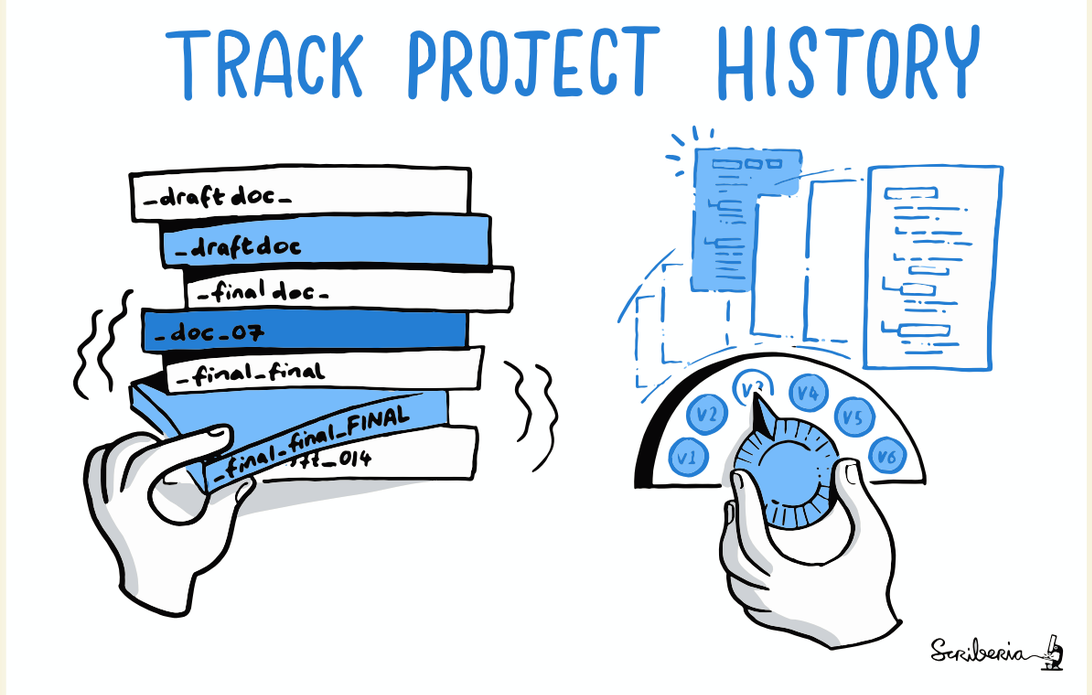
- keep things organized

WHAT IS VERSION CONTROL?



- keep things organized
- keep track of changes

WHAT IS VERSION CONTROL?



- keep things organized
- keep track of changes
- revert changes or go back to previous states

WHAT IS GIT? - A CRASH COURSE IN CODE

WHAT IS GIT? - A CRASH COURSE IN CODE

```
$ git init my_project # create a new git repo
Initialized empty Git repository in /Users/wittkuhn/Desktop/my_project/.git/
```

WHAT IS GIT? - A CRASH COURSE IN CODE

```
$ git init my_project # create a new git repo
Initialized empty Git repository in /Users/wittkuhn/Desktop/my_project/.git/
```

```
$ cd my_project # go into the my_project directory
```

WHAT IS GIT? - A CRASH COURSE IN CODE

```
$ git init my_project # create a new git repo
Initialized empty Git repository in /Users/wittkuhn/Desktop/my_project/.git/
```

```
$ cd my_project # go into the my_project directory
```

```
$ echo "hello world" >> README.md # create a README.md file
```

WHAT IS GIT? - A CRASH COURSE IN CODE

```
$ git init my_project # create a new git repo
Initialized empty Git repository in /Users/wittkuhn/Desktop/my_project/.git/
```

```
$ cd my_project # go into the my_project directory
```

```
$ echo "hello world" >> README.md # create a README.md file
```

```
$ ls # show the contents of the directory
README.md
```

WHAT IS GIT? - A CRASH COURSE IN CODE

```
$ git init my_project # create a new git repo
Initialized empty Git repository in /Users/wittkuhn/Desktop/my_project/.git/
```

```
$ cd my_project # go into the my_project directory
```

```
$ echo "hello world" >> README.md # create a README.md file
```

```
$ ls # show the contents of the directory
README.md
```

```
$ git add README.md # allow git to track changes in the README.md file
```

WHAT IS GIT? - A CRASH COURSE IN CODE

```
$ git init my_project # create a new git repo
Initialized empty Git repository in /Users/wittkuhn/Desktop/my_project/.git/
```

```
$ cd my_project # go into the my_project directory
```

```
$ echo "hello world" >> README.md # create a README.md file
```

```
$ ls # show the contents of the directory
README.md
```

```
$ git add README.md # allow git to track changes in the README.md file
```

```
$ git commit -m "initial commit" # commit the changes to your repo's history
[master (root-commit) 5118725] initial commit
1 file changed, 1 insertion(+)
create mode 100644 README.md
```

WHAT IS GIT? - A CRASH COURSE IN CODE

```
$ git init my_project # create a new git repo
Initialized empty Git repository in /Users/wittkuhn/Desktop/my_project/.git/
```

```
$ cd my_project # go into the my_project directory
```

```
$ echo "hello world" >> README.md # create a README.md file
```

```
$ ls # show the contents of the directory
README.md
```

```
$ git add README.md # allow git to track changes in the README.md file
```

```
$ git commit -m "initial commit" # commit the changes to your repo's history
[master (root-commit) 5118725] initial commit
1 file changed, 1 insertion(+)
create mode 100644 README.md
```

```
$ echo "goodbye world" >> README.md # add a new line to your README.md
$ git add README.md # allow git to track this recent change
$ git commit -m "update README.md" # commit the change to history
[master c56c4c0] update README.md
 1 file changed, 1 insertion(+)
```

WHAT IS GIT? - A CRASH COURSE IN CODE

```
$ git init my_project # create a new git repo
Initialized empty Git repository in /Users/wittkuhn/Desktop/my_project/.git/
```

```
$ cd my_project # go into the my_project directory
```

```
$ echo "hello world" >> README.md # create a README.md file
```

```
$ ls # show the contents of the directory
README.md
```

```
$ git add README.md # allow git to track changes in the README.md file
```

```
$ git commit -m "initial commit" # commit the changes to your repo's history
[master (root-commit) 5118725] initial commit
1 file changed, 1 insertion(+)
create mode 100644 README.md
```

```
$ echo "goodbye world" >> README.md # add a new line to your README.md
$ git add README.md # allow git to track this recent change
$ git commit -m "update README.md" # commit the change to history
[master c56c4c0] update README.md
 1 file changed, 1 insertion(+)
```

```
$ git log --oneline # show the commit history
c56c4c0 (HEAD -> master) update README.md
5118725 initial commit
```

WHAT IS GIT? - A CRASH COURSE IN CODE

```
$ git init my_project # create a new git repo
Initialized empty Git repository in /Users/wittkuhn/Desktop/my_project/.git/
```

```
$ cd my_project # go into the my_project directory
```

```
$ echo "hello world" >> README.md # create a README.md file
```

```
$ ls # show the contents of the directory
README.md
```

```
$ git add README.md # allow git to track changes in the README.md file
```

```
$ git commit -m "initial commit" # commit the changes to your repo's history
[master (root-commit) 5118725] initial commit
1 file changed, 1 insertion(+)
create mode 100644 README.md
```

```
$ echo "goodbye world" >> README.md # add a new line to your README.md
$ git add README.md # allow git to track this recent change
$ git commit -m "update README.md" # commit the change to history
[master c56c4c0] update README.md
 1 file changed, 1 insertion(+)
```

```
$ git log --oneline # show the commit history
c56c4c0 (HEAD -> master) update README.md
5118725 initial commit
```

```
$ git status
On branch master
nothing to commit, working tree clean
```

QUIZ QUESTION: WHAT IS INSIDE README.MD?

QUIZ QUESTION: WHAT IS INSIDE README.MD?

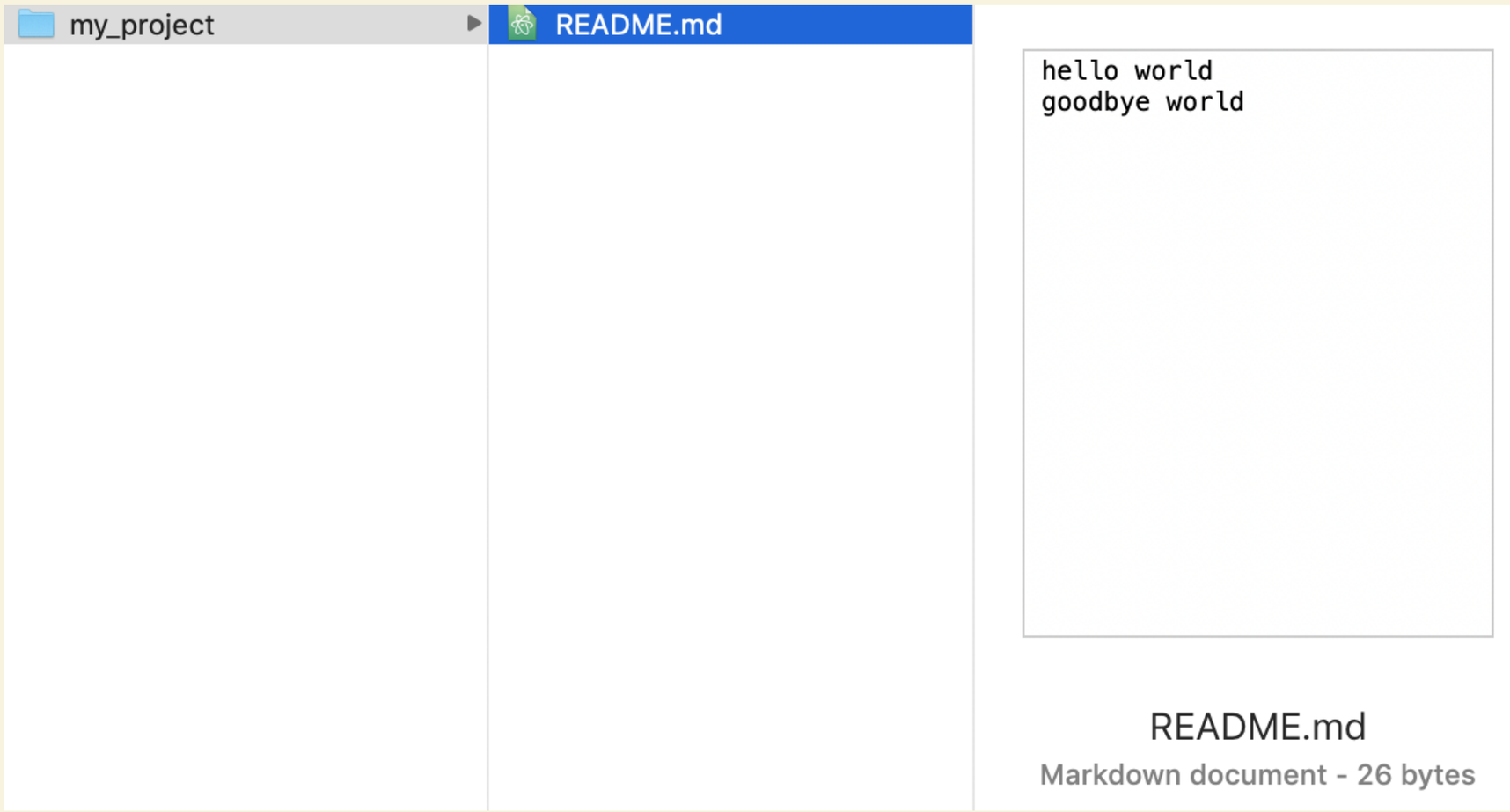
```
$ vim README.md

hello world
goodbye world
```

QUIZ QUESTION: WHAT IS INSIDE README.MD?

```
$ vim README.md

hello world
goodbye world
```



WHAT ARE GITHUB AND GITLAB?

GitHub and GitLab are platforms to host the contents of your Git repo

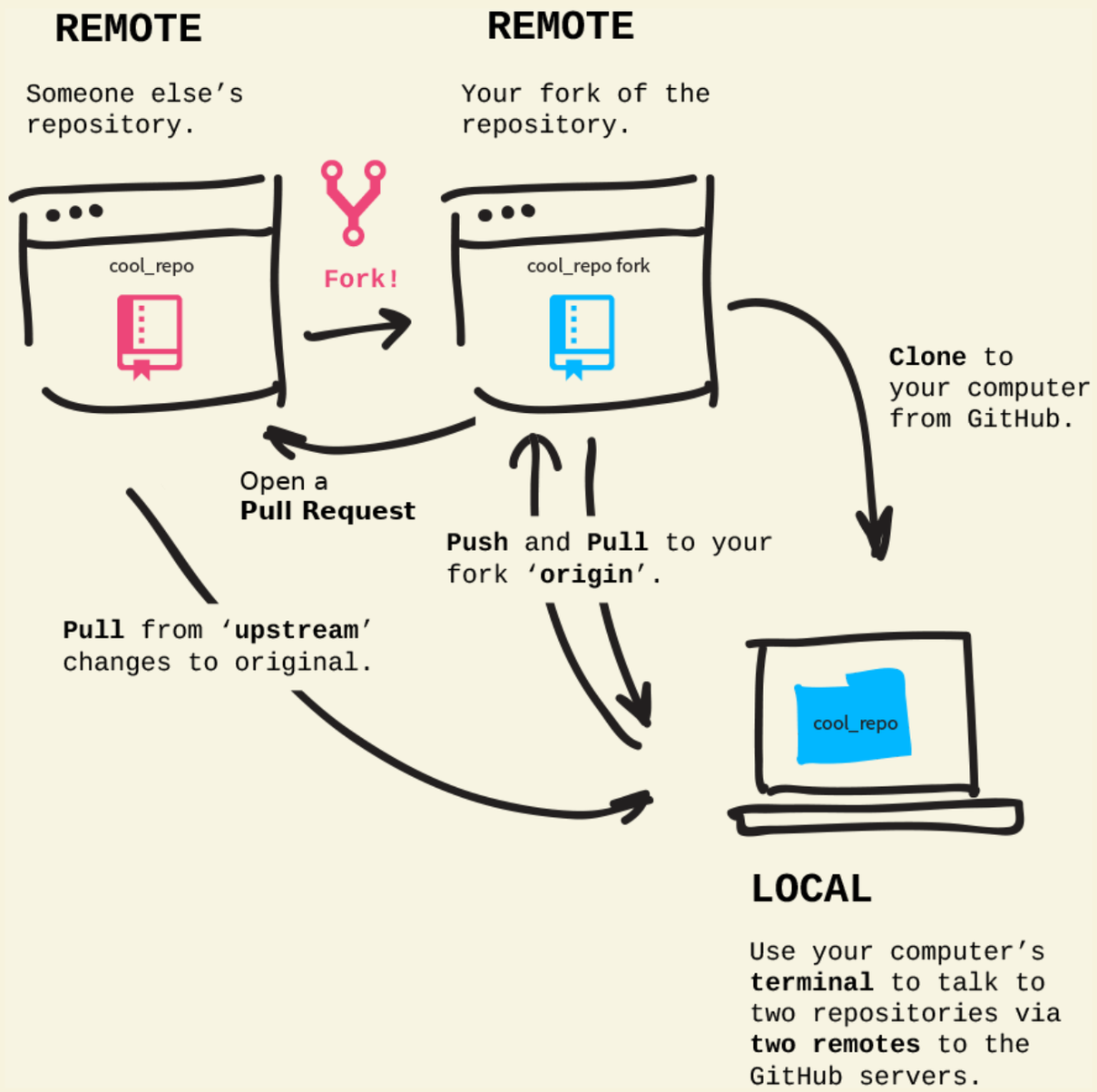


Image credit: https://docs.inm7.de/img/git_PR.png

We will present workflows that share data and code via these platforms

DataLad can help with small or large-scale data management

DataLad can help with small or large-scale data management

Free,
open source,
command line tool & Python API

- A command-line tool, available for all major operating systems (Linux, macOS/OSX, Windows), MIT-licensed
- Build on top of Git and Git-annex
- **Allows...**
  ... version-controlling arbitrarily large content
  version control data and software alongside to code!
  ... transport mechanisms for sharing and obtaining data
  consume and collaborate on data (analyses) like software
  ... (computationally) reproducible data analysis
  Track and share provenance of all digital objects
  ... and *much* more
- Completely domain-agnostic

ACKNOWLEDGEMENTS

Software

- Michael Hanke (INM-7)
- Yaroslav Halchenko
- Joey Hess (git-annex)
- Kyle Meyer
- Benjamin Poldrack (INM-7)
- *26 additional contributors*

Documentation project

- Michael Hanke (INM-7)
- Laura Waite (INM-7)
- *28 additional contributors*

Funders

NSF 1429999

germany-usa

SPONSORED BY THE

Federal Ministry of Education and Research

BMBF 01GQ1411

JÜLICH Forschungszentrum

EUROPEAN UNION
European Regional Development Fund

cbbs
center for behavioral brain sciences

SACHSEN-ANHALT
Ministerium für Wissenschaft und Wirtschaft

Collaborators

Human Brain Project

CONP PCNO

VirtualBrainCloud

OpenNEURO

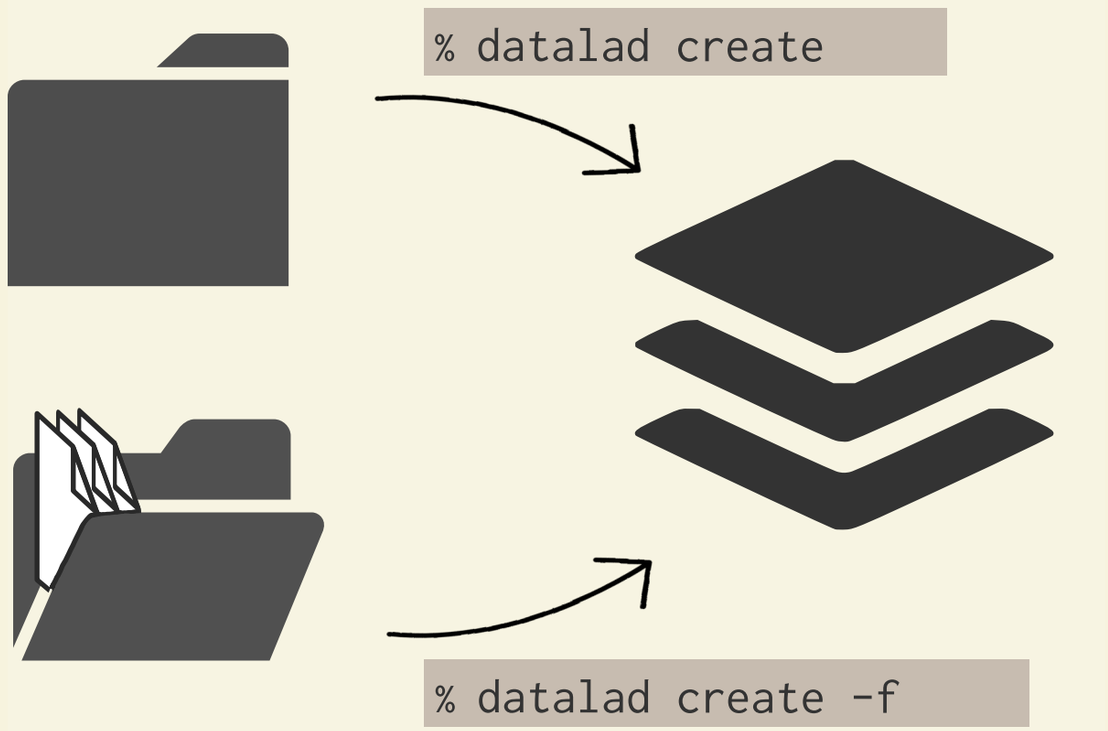
cbrain

brainlife.io

CORE CONCEPTS & FEATURES

EVERYTHING HAPPENS IN DATALAD DATASETS

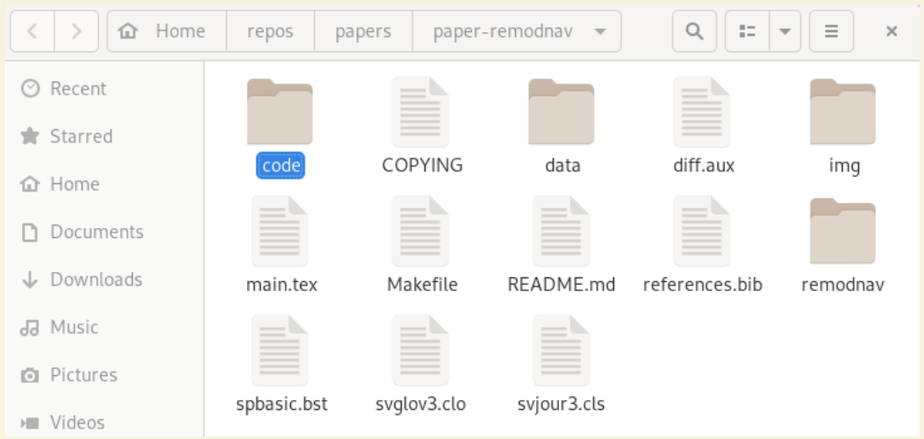
create new, empty datasets to populate...

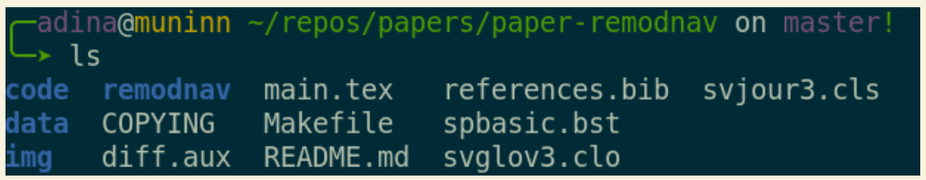


.... or transform existing directories into datasets

DATASET = GIT/GIT-ANNEX REPOSITORY

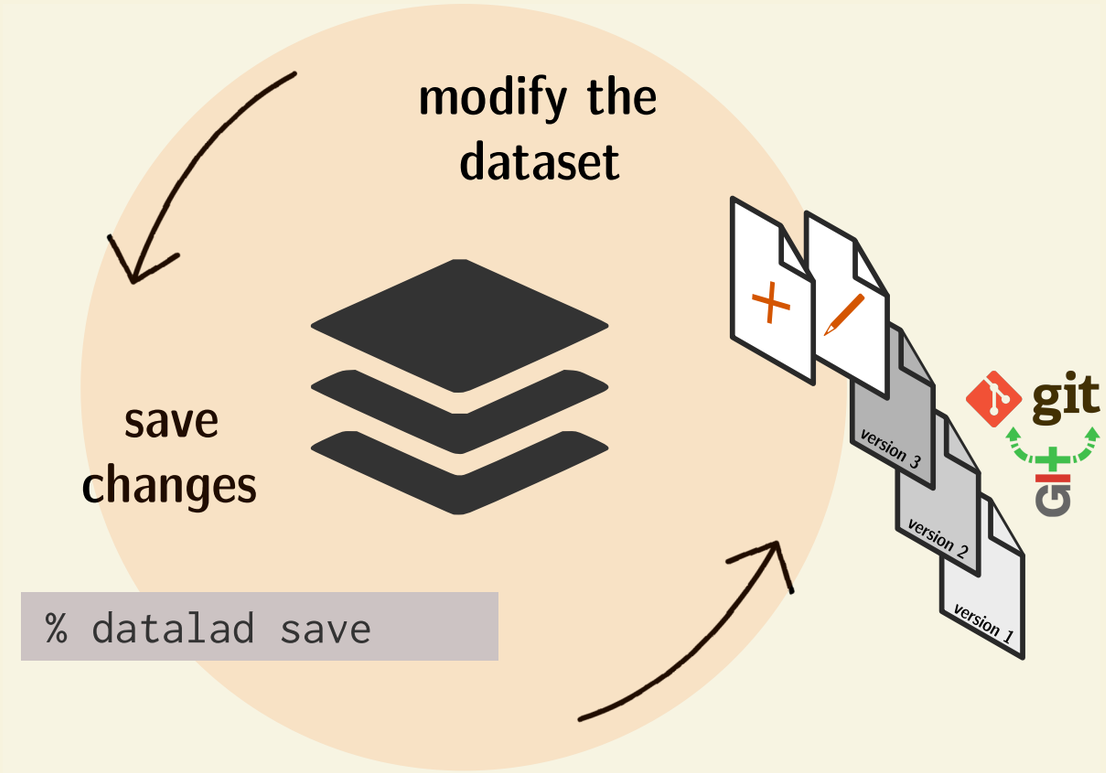
- content agnostic
- no custom data structures
- complete decentralization
- Looks and feels like a directory on your computer:



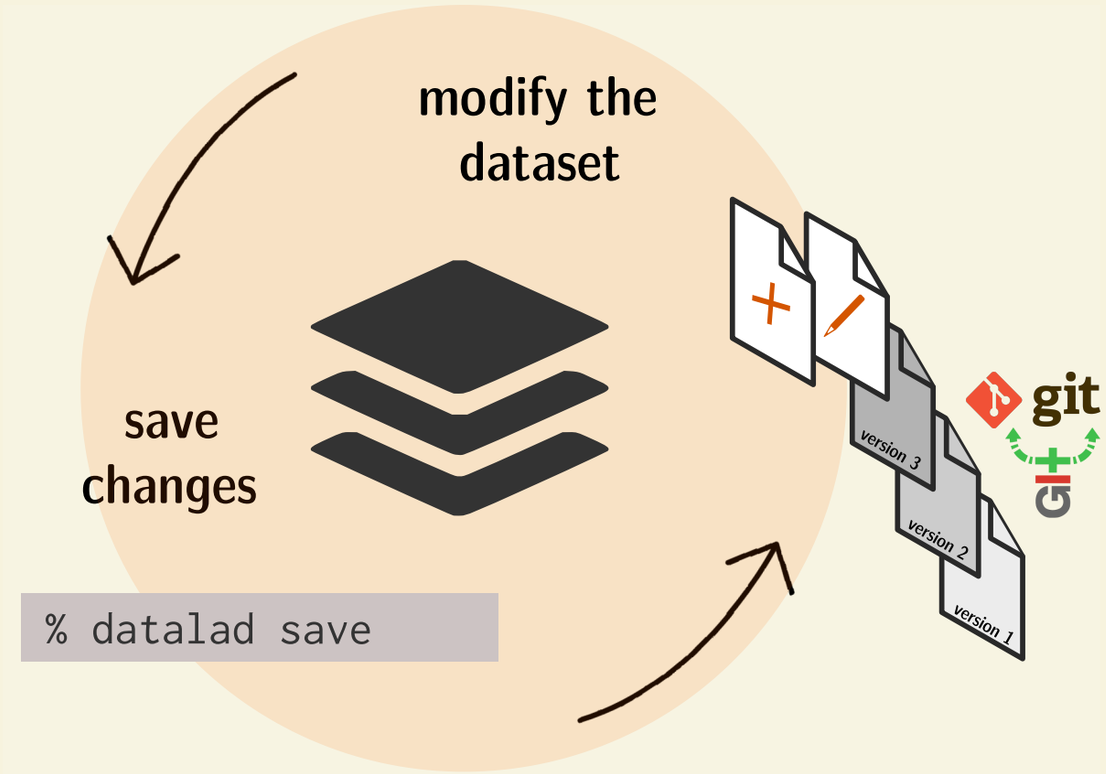


File viewer and terminal view of a DataLad dataset

VERSION CONTROL ARBITRARILY LARGE FILES

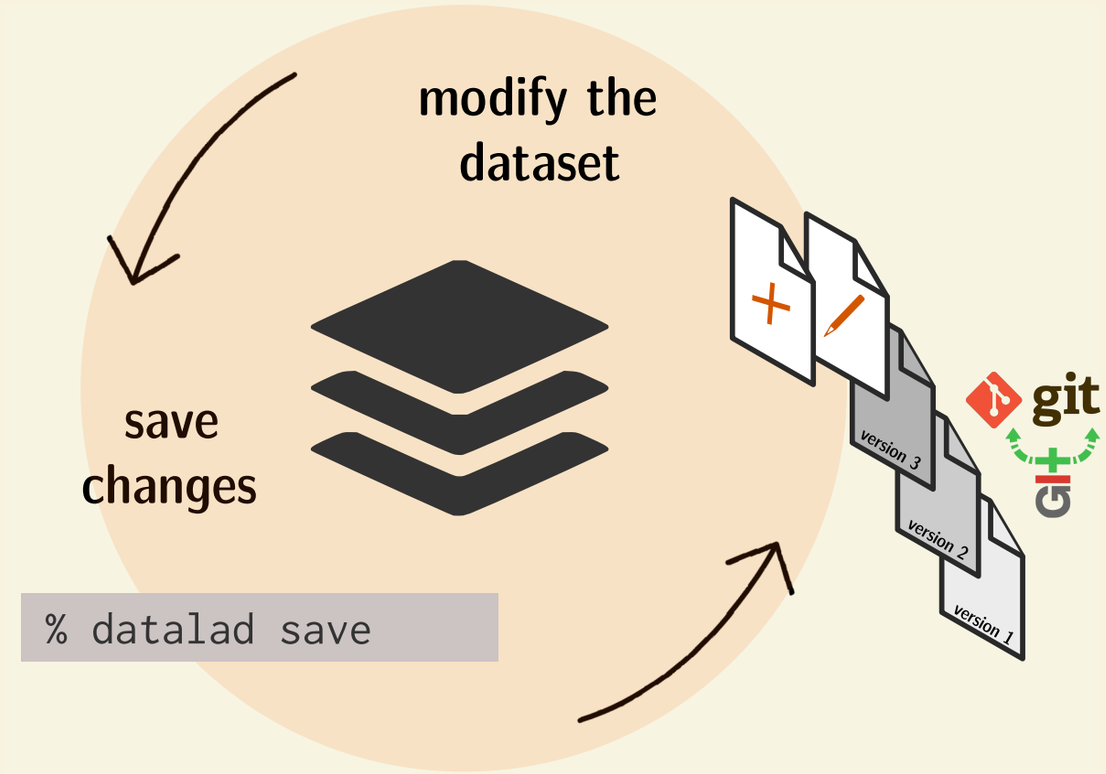


VERSION CONTROL ARBITRARILY LARGE FILES



Stay flexible:

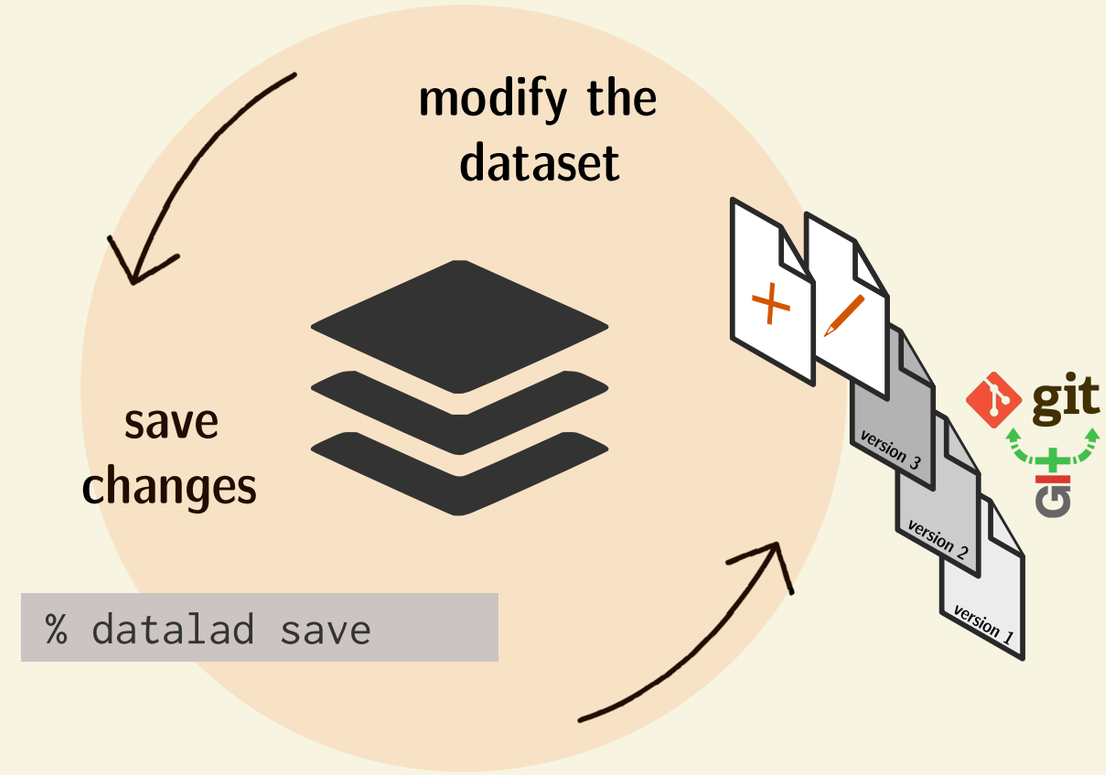
VERSION CONTROL ARBITRARILY LARGE FILES



Stay flexible:

- Non-complex DataLad core API (easy for data management novices)

VERSION CONTROL ARBITRARILY LARGE FILES



Stay flexible:

- Non-complex DataLad core API (easy for data management novices)
- Pure Git or git-annex commands (for regular Git or git-annex users, or to use specific functionality)

USE A DATASETS' HISTORY

```
2020-06-05 10:58 +0200 Adina Wagner  M─┐ [master] {upstream/master} {upstream/HEAD} Merge pull request #12 from psychoinformatics-d
2020-06-05 08:24 +0200 Adina Wagner  │ o [finalround] {upstream/finalround} add results from computing with mean instead of median
2020-06-05 09:09 +0200 Michael Hanke o │ Change wording, clarify comment
2020-06-05 07:26 +0200 Michael Hanke M─┼─┐ Merge remote-tracking branch 'gh-mine/finalround'
2020-05-28 16:39 +0200 Asim H Dar    │ o─┘ Added datalad.get() so S2SRMS() pulls data and can run standalone
2020-05-18 08:25 +0200 Adina Wagner  │ o {gh-asim/finalround} S2SRMS: example implementation of the S2SRMS method suggested by R2
2020-05-01 17:38 +0200 Adina Wagner  │ o Minor edits as suggested by reviewer 2
2020-05-29 09:04 +0200 Adina Wagner  M─┼─┐ Merge pull request #13 from psychoinformatics-de/adswa-patch-1
2020-05-24 09:15 +0200 Adina Wagner  │ │ o {upstream/adswa-patch-1} Fix installation instructions
2020-05-24 09:53 +0200 Adina Wagner  M─┼─┼─┐ Merge pull request #14 from psychoinformatics-de/bf-data
2020-05-24 09:33 +0200 Adina Wagner  │ │ │ o [bf-data] One-time datalad import
2020-05-24 09:32 +0200 Adina Wagner  │ │ │ o install and get relevant subdataset data
2020-03-18 10:19 +0100 Michael Hanke M─┼─┼─┘ Merge pull request #8 from psychoinformatics-de/adswa-patch-1
2019-12-19 10:22 +0100 Adina Wagner  │ o {gh-asim/adswa-patch-1} add sklearn to requirements
2020-03-18 10:13 +0100 Michael Hanke o │ Tune new figure caption
2020-03-18 10:03 +0100 Michael Hanke M─┼─┐ Merge pull request #11 from ElectronicTeaCup/revision_2
2020-03-18 09:59 +0100 Adina Wagner  │ M─┼─┐ [revision_2] {gh-asim/revision_2} Merge branch 'revision_2' of github.com:ElectronicTe
2020-03-18 09:58 +0100 Michael Hanke │ │ o │ Last detections
2020-03-18 09:59 +0100 Adina Wagner  │ │ o │ name parameter in caption
```

Date Author Change summary (commit message)

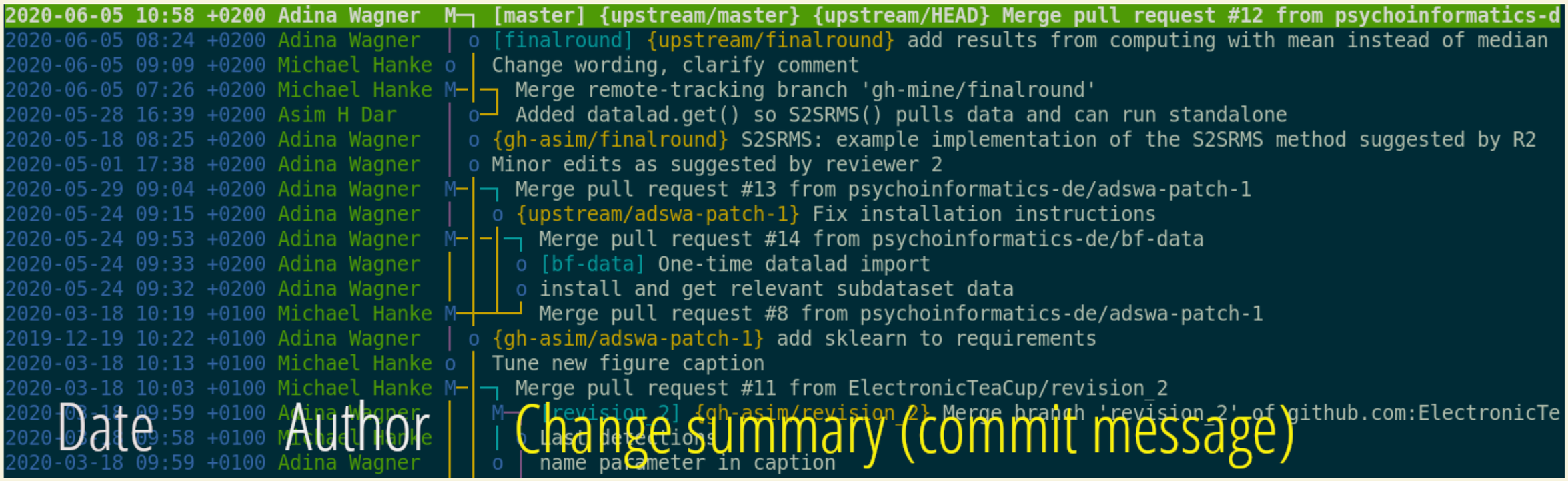
USE A DATASETS' HISTORY

```
2020-06-05 10:58 +0200 Adina Wagner  M─┐ [master] {upstream/master} {upstream/HEAD} Merge pull request #12 from psychoinformatics-d
2020-06-05 08:24 +0200 Adina Wagner  │ o [finalround] {upstream/finalround} add results from computing with mean instead of median
2020-06-05 09:09 +0200 Michael Hanke o │ Change wording, clarify comment
2020-06-05 07:26 +0200 Michael Hanke M─┼─┐ Merge remote-tracking branch 'gh-mine/finalround'
2020-05-28 16:39 +0200 Asim H Dar    │ o─┘ Added datalad.get() so S2SRMS() pulls data and can run standalone
2020-05-18 08:25 +0200 Adina Wagner  │ o {gh-asim/finalround} S2SRMS: example implementation of the S2SRMS method suggested by R2
2020-05-01 17:38 +0200 Adina Wagner  │ o Minor edits as suggested by reviewer 2
2020-05-29 09:04 +0200 Adina Wagner  M─┼─┐ Merge pull request #13 from psychoinformatics-de/adswa-patch-1
2020-05-24 09:15 +0200 Adina Wagner  │ │ o {upstream/adswa-patch-1} Fix installation instructions
2020-05-24 09:53 +0200 Adina Wagner  M─┼─┼─┐ Merge pull request #14 from psychoinformatics-de/bf-data
2020-05-24 09:33 +0200 Adina Wagner  │ │ │ o [bf-data] One-time datalad import
2020-05-24 09:32 +0200 Adina Wagner  │ │ │ o install and get relevant subdataset data
2020-03-18 10:19 +0100 Michael Hanke M─┼─┼─┘ Merge pull request #8 from psychoinformatics-de/adswa-patch-1
2019-12-19 10:22 +0100 Adina Wagner  │ o {gh-asim/adswa-patch-1} add sklearn to requirements
2020-03-18 10:13 +0100 Michael Hanke o │ Tune new figure caption
2020-03-18 10:03 +0100 Michael Hanke M─┼─┐ Merge pull request #11 from ElectronicTeaCup/revision_2
2020-03-18 09:59 +0100 Adina Wagner  │ M─┼─┐ [revision_2] {gh-asim/revision_2} Merge branch 'revision_2' of github.com:ElectronicTe
2020-03-18 09:58 +0100 Michael Hanke │ │ o │ Last detections
2020-03-18 09:59 +0100 Adina Wagner  │ │ o │ name parameter in caption
```

Date Author Change summary (commit message)

- reset your dataset (or subset of it) to a previous state,

USE A DATASETS' HISTORY



- reset your dataset (or subset of it) to a previous state,
- revert changes or bring them back,

USE A DATASETS' HISTORY

```
2020-06-05 10:58 +0200 Adina Wagner  M─┐ [master] {upstream/master} {upstream/HEAD} Merge pull request #12 from psychoinformatics-d
2020-06-05 08:24 +0200 Adina Wagner  │ o [finalround] {upstream/finalround} add results from computing with mean instead of median
2020-06-05 09:09 +0200 Michael Hanke o │ Change wording, clarify comment
2020-06-05 07:26 +0200 Michael Hanke M─┼─┐ Merge remote-tracking branch 'gh-mine/finalround'
2020-05-28 16:39 +0200 Asim H Dar    │ o─┘ Added datalad.get() so S2SRMS() pulls data and can run standalone
2020-05-18 08:25 +0200 Adina Wagner  │ o {gh-asim/finalround} S2SRMS: example implementation of the S2SRMS method suggested by R2
2020-05-01 17:38 +0200 Adina Wagner  │ o Minor edits as suggested by reviewer 2
2020-05-29 09:04 +0200 Adina Wagner  M─┼─┐ Merge pull request #13 from psychoinformatics-de/adswa-patch-1
2020-05-24 09:15 +0200 Adina Wagner  │ │ o {upstream/adswa-patch-1} Fix installation instructions
2020-05-24 09:53 +0200 Adina Wagner  M─┼─┼─┐ Merge pull request #14 from psychoinformatics-de/bf-data
2020-05-24 09:33 +0200 Adina Wagner  │ │ │ o [bf-data] One-time datalad import
2020-05-24 09:32 +0200 Adina Wagner  │ │ │ o install and get relevant subdataset data
2020-03-18 10:19 +0100 Michael Hanke M─┼─┼─┘ Merge pull request #8 from psychoinformatics-de/adswa-patch-1
2019-12-19 10:22 +0100 Adina Wagner  │ o {gh-asim/adswa-patch-1} add sklearn to requirements
2020-03-18 10:13 +0100 Michael Hanke o │ Tune new figure caption
2020-03-18 10:03 +0100 Michael Hanke M─┼─┐ Merge pull request #11 from ElectronicTeaCup/revision_2
2020-03-18 09:59 +0100 Adina Wagner  │ M─┼─┐ [revision_2] {gh-asim/revision_2} Merge branch 'revision_2' of github.com:ElectronicTe
2020-03-18 09:58 +0100 Michael Hanke │ │ o │ Last detections
2020-03-18 09:59 +0100 Adina Wagner  │ │ o │ name parameter in caption
```

Date Author Change summary (commit message)

- reset your dataset (or subset of it) to a previous state,
- revert changes or bring them back,
- find out what was done when, how, why, and by whom

USE A DATASETS' HISTORY

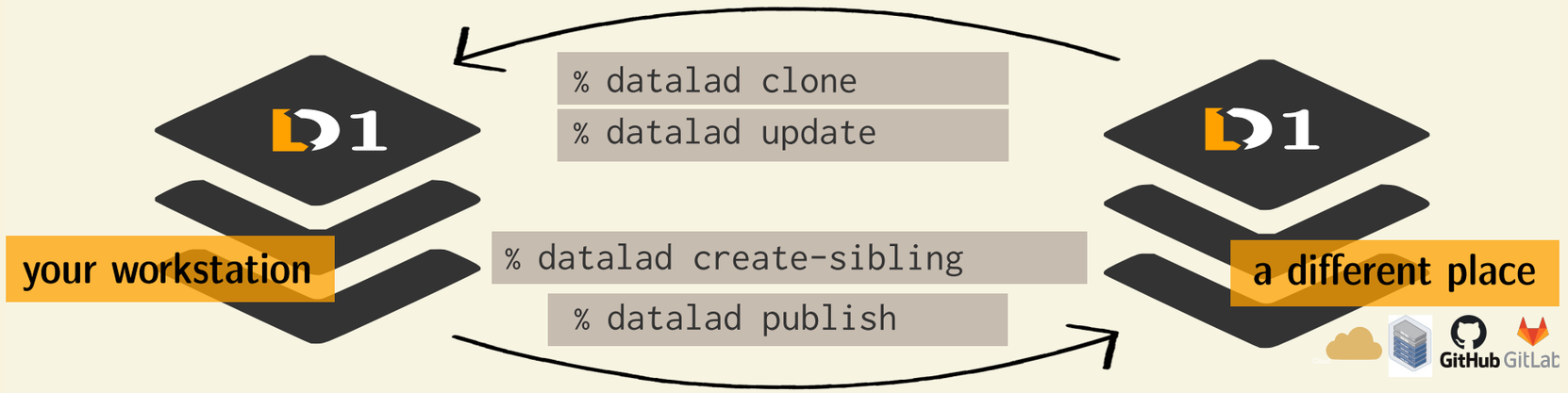
```
2020-06-05 10:58 +0200 Adina Wagner  M─┐ [master] {upstream/master} {upstream/HEAD} Merge pull request #12 from psychoinformatics-d
2020-06-05 08:24 +0200 Adina Wagner  │ o [finalround] {upstream/finalround} add results from computing with mean instead of median
2020-06-05 09:09 +0200 Michael Hanke o │ Change wording, clarify comment
2020-06-05 07:26 +0200 Michael Hanke M─┼─┐ Merge remote-tracking branch 'gh-mine/finalround'
2020-05-28 16:39 +0200 Asim H Dar    │ o─┘ Added datalad.get() so S2SRMS() pulls data and can run standalone
2020-05-18 08:25 +0200 Adina Wagner  │ o {gh-asim/finalround} S2SRMS: example implementation of the S2SRMS method suggested by R2
2020-05-01 17:38 +0200 Adina Wagner  │ o Minor edits as suggested by reviewer 2
2020-05-29 09:04 +0200 Adina Wagner  M─┼─┐ Merge pull request #13 from psychoinformatics-de/adswa-patch-1
2020-05-24 09:15 +0200 Adina Wagner  │ │ o {upstream/adswa-patch-1} Fix installation instructions
2020-05-24 09:53 +0200 Adina Wagner  M─┼─┼─┐ Merge pull request #14 from psychoinformatics-de/bf-data
2020-05-24 09:33 +0200 Adina Wagner  │ │ │ o [bf-data] One-time datalad import
2020-05-24 09:32 +0200 Adina Wagner  │ │ │ o install and get relevant subdataset data
2020-03-18 10:19 +0100 Michael Hanke M─┼─┼─┘ Merge pull request #8 from psychoinformatics-de/adswa-patch-1
2019-12-19 10:22 +0100 Adina Wagner  │ o {gh-asim/adswa-patch-1} add sklearn to requirements
2020-03-18 10:13 +0100 Michael Hanke o │ Tune new figure caption
2020-03-18 10:03 +0100 Michael Hanke M─┼─┐ Merge pull request #11 from ElectronicTeaCup/revision_2
2020-03-18 09:59 +0100 Adina Wagner  │ M─┼─┐ {revision_2] {gh-asim/revision_2} Merge branch 'revision_2' of github.com:ElectronicTe
2020-03-18 09:58 +0100 Michael Hanke │ │ o │ Last detections
2020-03-18 09:59 +0100 Adina Wagner  │ │ o │ name parameter in caption
```

Date Author Change summary (commit message)

- reset your dataset (or subset of it) to a previous state,
- revert changes or bring them back,
- find out what was done when, how, why, and by whom
- Identify precise versions: Use data in the most recent version, or the one from 2018, or...

CONSUME AND COLLABORATE

Consume existing datasets and stay up-to-date

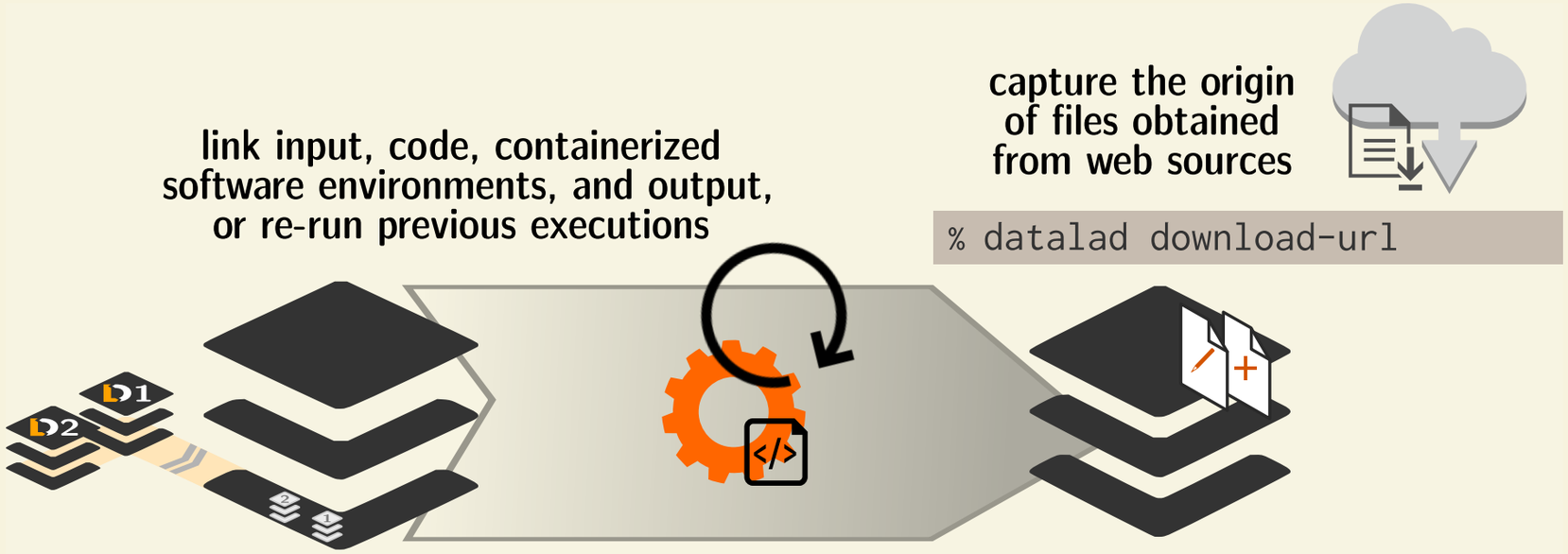


Create sibling datasets to publish to or update from

MACHINE-READABLE, RE-EXECUTABLE PROVENANCE

link input, code, containerized
software environments, and output,
or re-run previous executions

capture the origin
of files obtained
from web sources



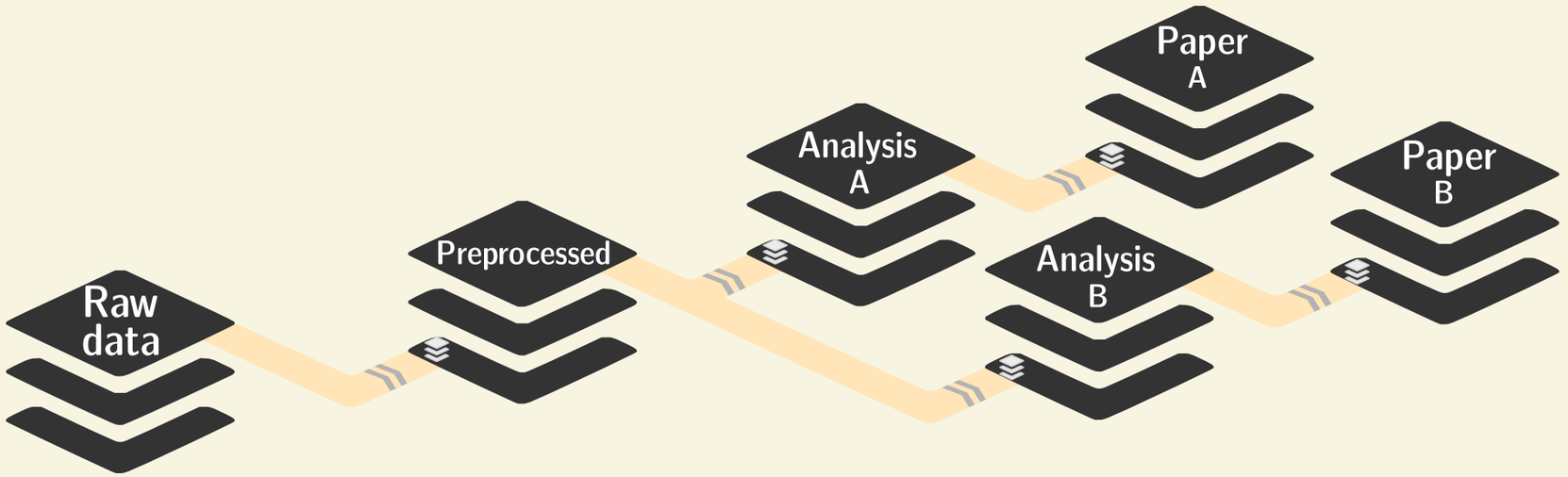

```
% datalad download-url
```

```
% datalad run
```

```
% datalad rerun
```

```
% datalad run-procedure
```

SEAMLESS NESTING AND DATASET LINKAGE



Nest modular datasets to create a linked hierarchy of datasets,
and enable recursive operations throughout the hierarchy

THIRD PARTY INTEGRATIONS

Apart from **local computing infrastructure** (from private laptops to computational clusters), datasets can be hosted in major **third party repository hosting and cloud storage** services. More info: Chapter on Third party infrastructure.

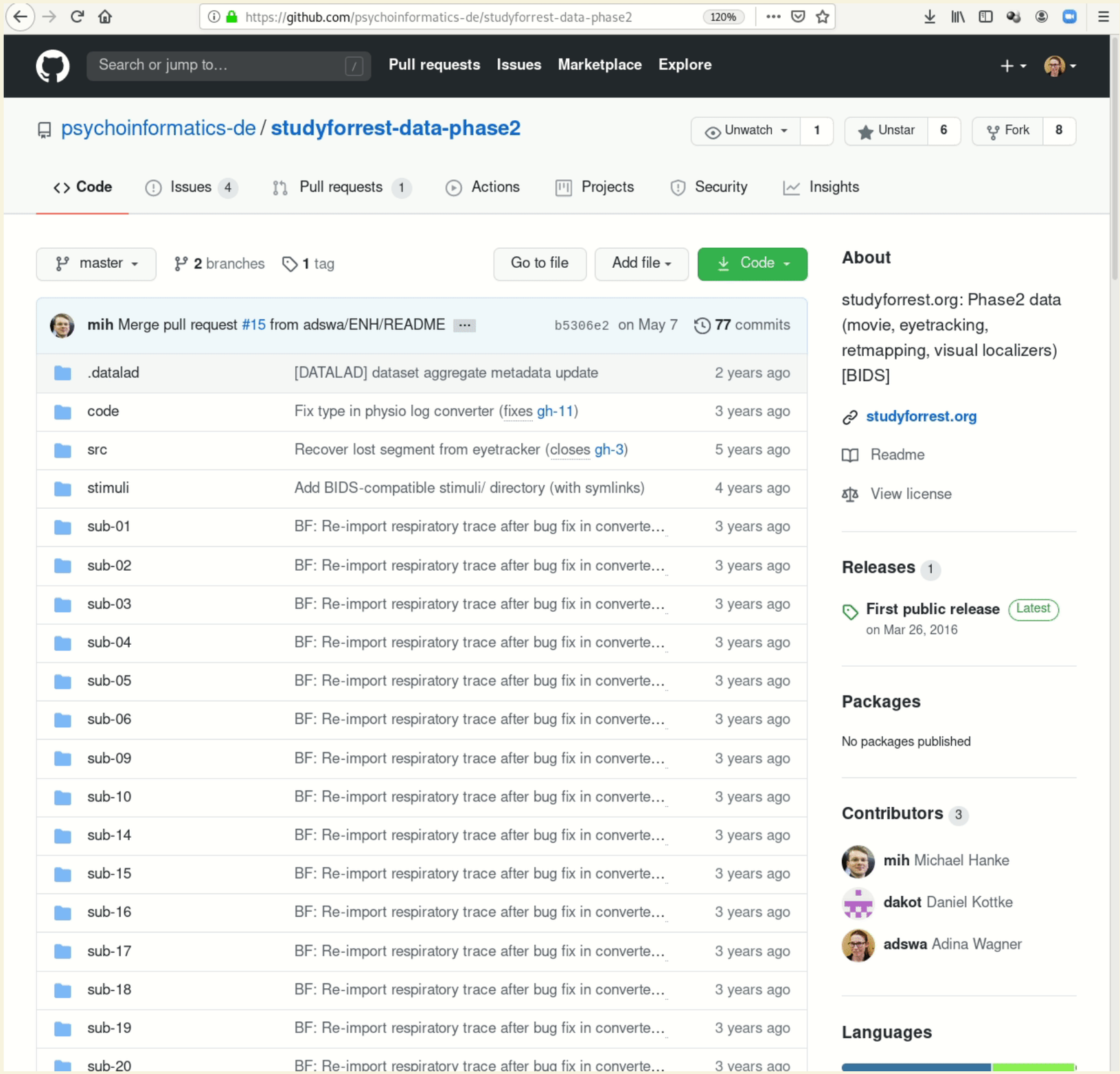
EXAMPLES OF WHAT DATALAD CAN BE USED FOR:

EXAMPLES OF WHAT DATALAD CAN BE USED FOR:

- **Publish or consume datasets** via GitHub, GitLab, OSF, or similar services

EXAMPLES OF WHAT DATALAD CAN BE USED FOR:

- **Publish or consume datasets** via GitHub, GitLab, OSF, or similar services



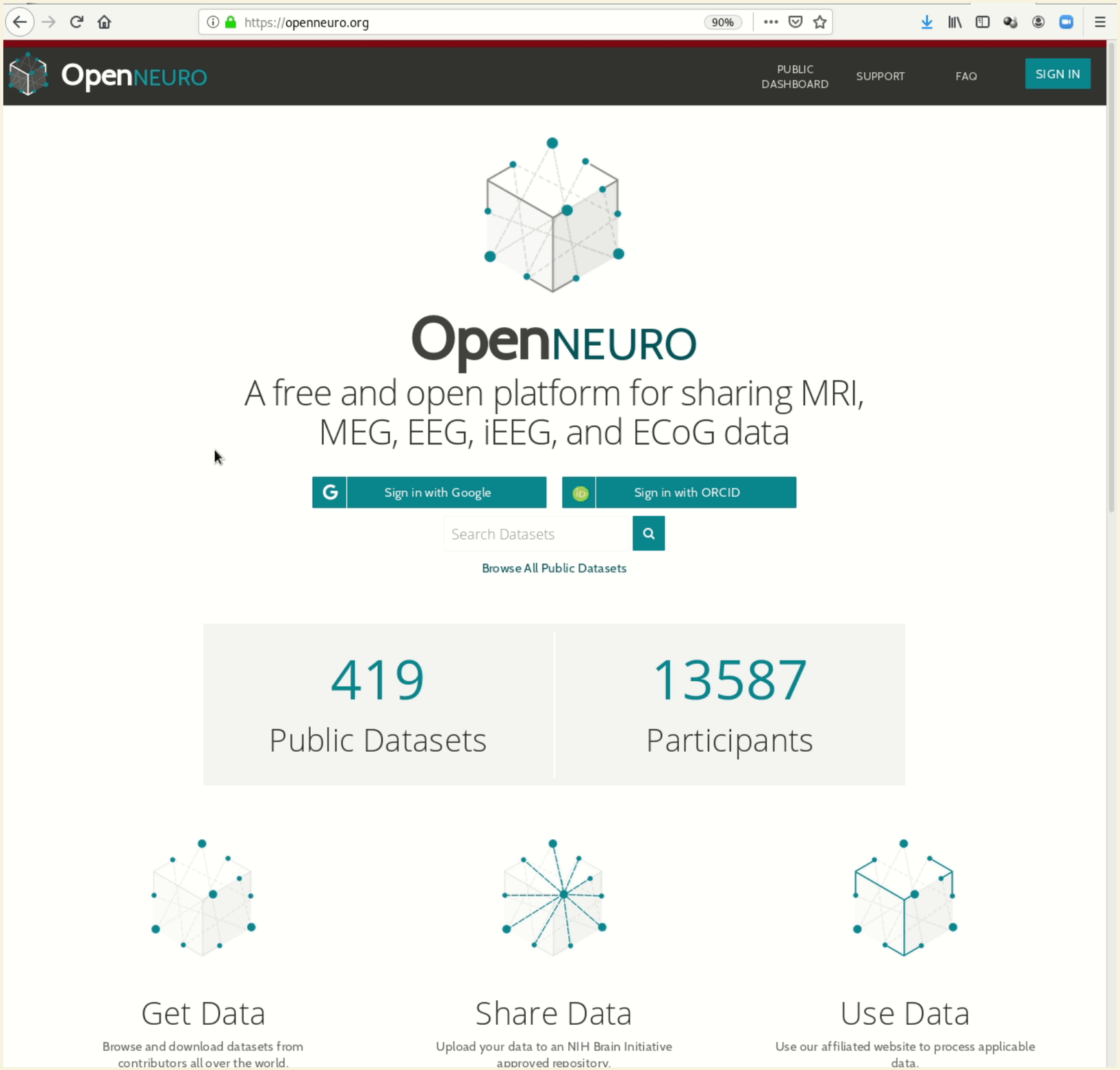
EXAMPLES OF WHAT DATALAD CAN BE USED FOR:

EXAMPLES OF WHAT DATALAD CAN BE USED FOR:

- Behind-the-scenes **infrastructure component for data transport and versioning** (e.g., used by OpenNeuro, brainlife.io , the Canadian Open Neuroscience Platform (CONP), CBRAIN)

EXAMPLES OF WHAT DATALAD CAN BE USED FOR:

- Behind-the-scenes **infrastructure component for data transport and versioning** (e.g., used by OpenNeuro, brainlife.io , the Canadian Open Neuroscience Platform (CONP), CBRAIN)



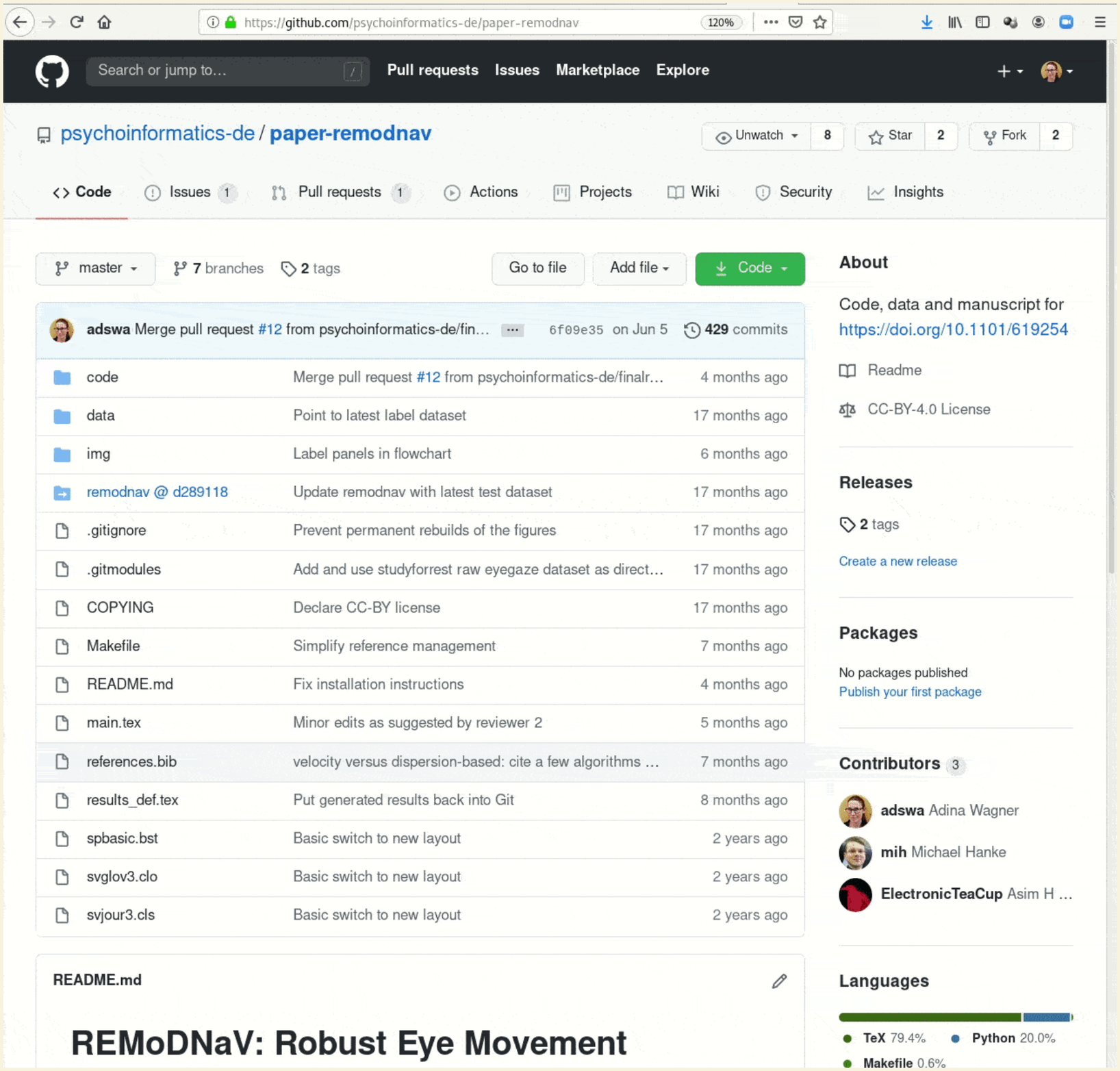
EXAMPLES OF WHAT DATALAD CAN BE USED FOR:

EXAMPLES OF WHAT DATALAD CAN BE USED FOR:

- **Creating and sharing reproducible, open science**: Sharing data, software, code, and provenance

EXAMPLES OF WHAT DATALAD CAN BE USED FOR:

- **Creating and sharing reproducible, open science**: Sharing data, software, code, and provenance



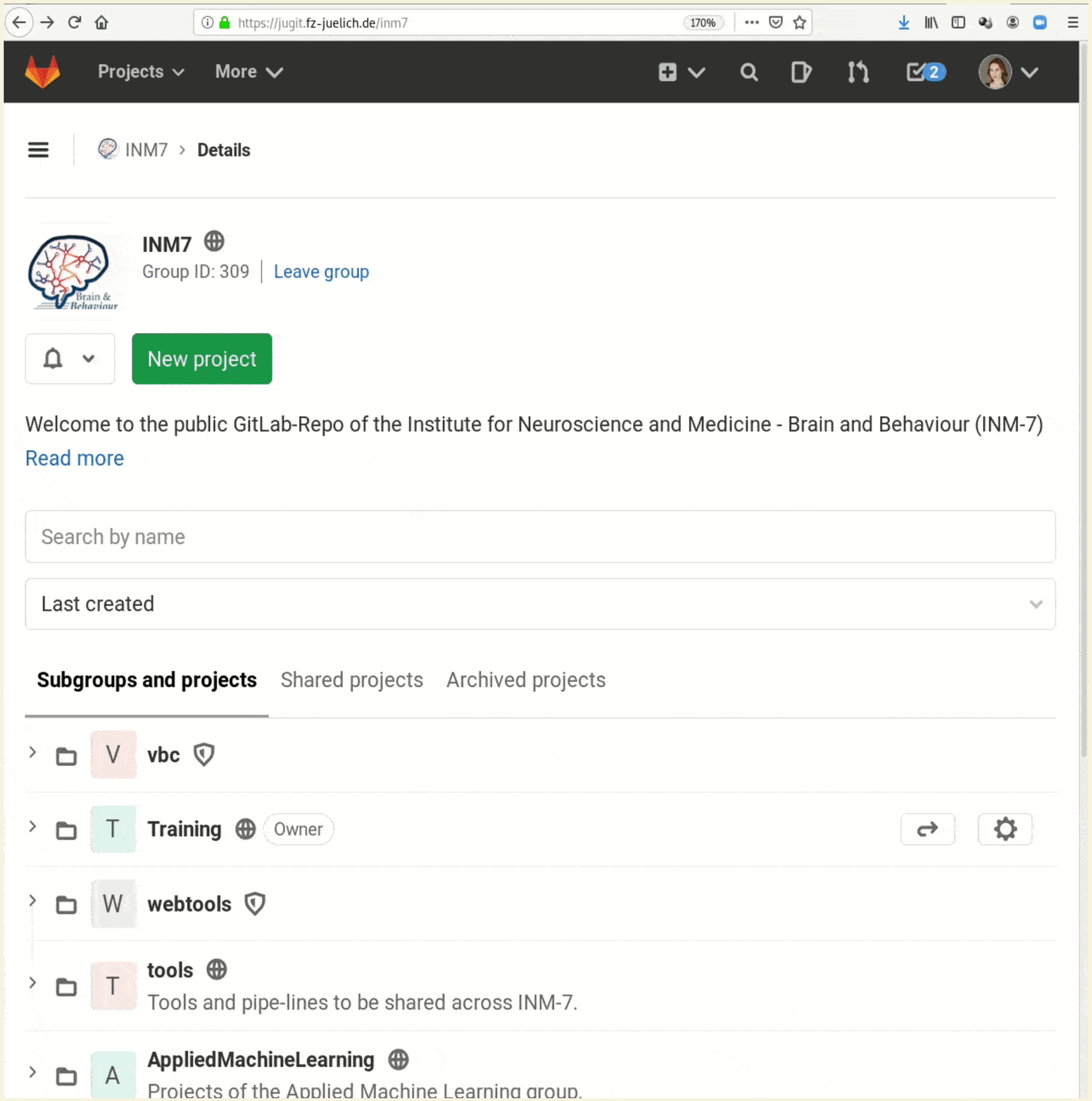
EXAMPLES OF WHAT DATALAD CAN BE USED FOR:

EXAMPLES OF WHAT DATALAD CAN BE USED FOR:

- **Central data management** and archival system

EXAMPLES OF WHAT DATALAD CAN BE USED FOR:

- **Central data management** and archival system



... AND MANY MORE!

Let's have a ☕, and then get started