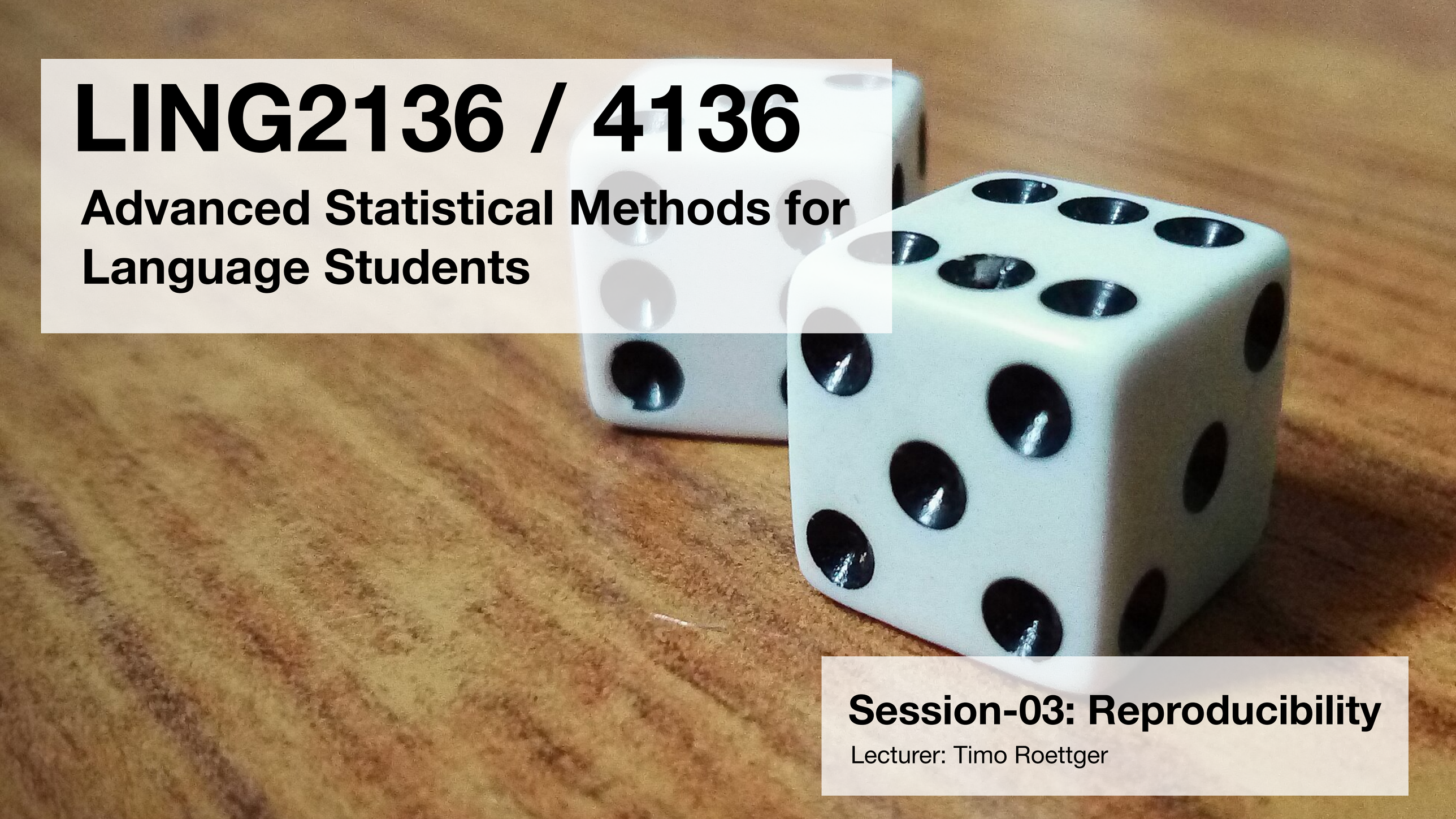


LING2136 / 4136

Advanced Statistical Methods for Language Students

Session-03: Reproducibility

Lecturer: Timo Roettger

file/exercises/
03_exercise.pdf

REPRODUCIBILITY

Empirical R.

Computational R.

Statistical R.

REPRODUCIBILITY

Empirical R.

e.g. provide detailed information about the data collection procedure and (ideally) provide the raw data.

Computational R.

Statistical R.

REPRODUCIBILITY

Empirical R. e.g. provide detailed information about the data collection procedure and (ideally) provide the raw data.

Computational R. **e.g. provide detailed information about the software (version) used and provide the code (and data table) that generated summary statistics, plots, inferential estimates, etc.**

Statistical R.

REPRODUCIBILITY

Empirical R. e.g. provide detailed information about the data collection procedure and (ideally) provide the raw data.

Computational R. e.g. provide detailed information about the software (version) used and provide the code (and data table) that generated summary statistics, plots, inferential estimates, etc.

Statistical R. **e.g. provide detailed information about the choice of statistical models, model parameters, number of tests performed. Ideally this information should be provided before data has been observed (= preregistration)**

REPRODUCIBILITY

Empirical R.

e.g. provide detailed information about the data collection procedure and (ideally) provide the **raw data**.

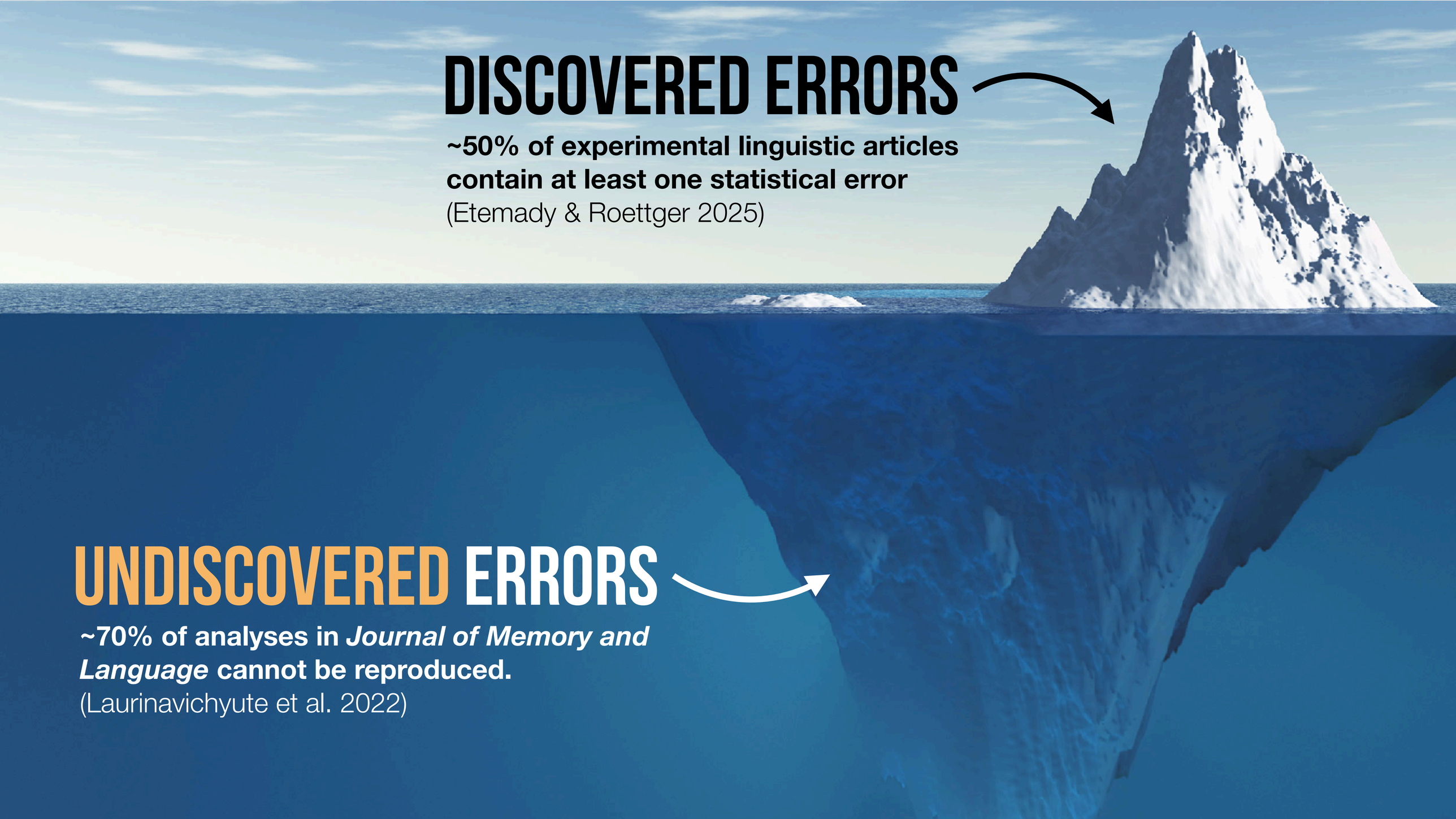
Computational R.

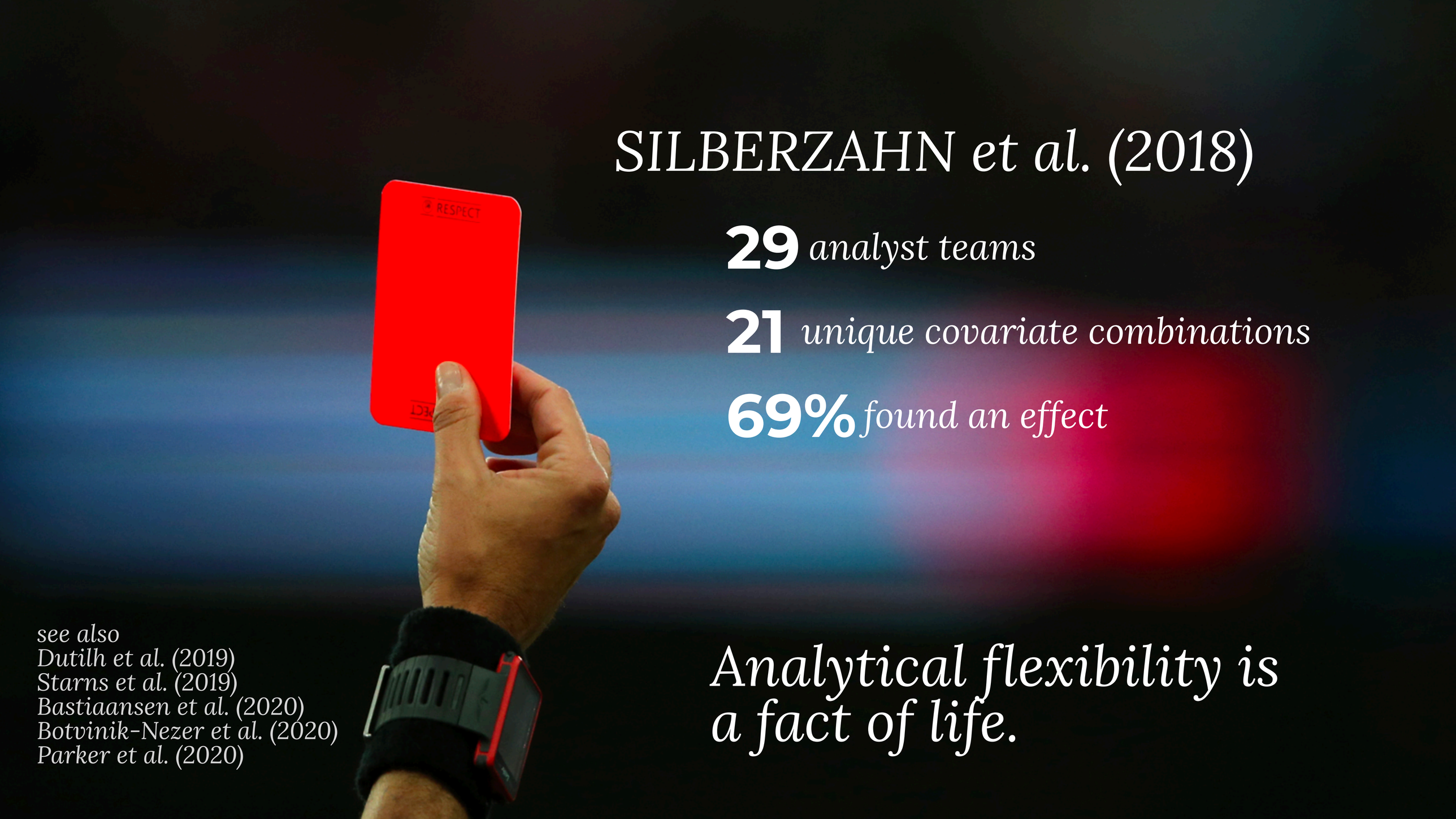
e.g. provide detailed information about the software (version) used and provide the **code** (and **data table**) that generated summary statistics, plots, inferential estimates, etc.

Statistical R.

e.g. provide detailed information about the choice of **statistical models, model parameters, number of tests** performed. Ideally this information should be provided before data has been observed (> preregistration)


68% OF DATA
CANNOT BE
ACCESSED
Hardwicke & Ioannidis (2018). *PloS one*.
BURG
WÄCHTER
KARAT


DISCOVERED ERRORS
~50% of experimental linguistic articles contain at least one statistical error
(Etemady & Roettger 2025)
UNDISCOVERED ERRORS
~70% of analyses in Journal of Memory and Language cannot be reproduced.
(Laurinavichyute et al. 2022)



SILBERZAHN *et al.* (2018)

29 *analyst teams*

21 *unique covariate combinations*

69% *found an effect*

see also
Dutilh et al. (2019)
Starns et al. (2019)
Bastiaansen et al. (2020)
Botvinik-Nezer et al. (2020)
Parker et al. (2020)

Analytical flexibility is a fact of life.

OPEN DATA: WHAT'S IN IT FOR YOU?

Piwowar & Vision (2013), McKiernan et al. (2016),
Steegen et al. (2016), Klein et al. (2018)

- **error detection** before publication
- **protects** against data loss
- increases **visibility** and facilitates access to **collaboration**, **jobs**, and **funding**
- **citation benefit**
- **saves resources** for future follow-ups
- enables **evidence accumulation**
- enables **critical evaluation**
- (easier peer review process)

“I don’t have time…”

(Borgman, 2012; Houtkoop et al. 2018)

“I am worried of getting scooped”

(Houtkoop et al. 2018)

```
102             mean_acc = mean(acc, na.rm = T),
103             mean_dev_ideal = mean(dev_ideal, na.rm
104
105
106 ## load in picture for screen background
107 image <- png::readPNG("/Users/timoroet       /mouse_tracked_UNreliabl       intonation/RF1/screen_dummy.png")
108
109 RF1_xpos <-
110   ggplot(xagg, aes(x = mean_time, y                , colour = Focus, fill = Focus       shape = Group)) +
111   geom_segment(x = c(332), y = -Inf                32), yend = Inf, lty = "dotted       size = 1, colour = "grey") +
112   geom_segment(x = c(700), y = -Inf                00), yend = Inf, lty = "dotted       size = 1, colour = "grey") +
113   annotate("text", x = 161, y = -1,                 Wuggy", size = 8, hjust = 0       vjust = 0.5) +
114   annotate("text", x = 516, y = -1,                dann die", size = 8, hjust       .5, vjust = 0.5) +
115   annotate("text", x = 750, y = -1,                RENT]...", size = 8, hj       0, vjust = 0.5) +
116   geom_path() +
117   geom_point(size = 2) +
118
```

“I don’t want others to find errors in my scripts”

(Houtkoop et al. 2018)

OPEN DATA: WHAT'S IN IT FOR YOU?

Piwowar & Vision (2013), McKiernan et al. (2016),
Steegen et al. (2016), Klein et al. (2018)

- **error detection** before publication
- **protects** against data loss
- increases **visibility** and facilitates access to **collaboration**, **jobs**, and **funding**
- **citation benefit**
- **saves resources** for future follow-ups
- enables **evidence accumulation**
- enables **critical evaluation**
- (easier peer review process)

HOW TO BE REPRODUCIBLE?

F

Findable

A

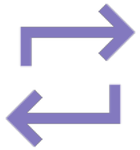
Accessible

I

Interoperable

R

Reusable



F

Findable

Store data and analysis in a publicly available place which ideally

uses persistent, unique identifiers (DOIs)

allows to add structured metadata

tracks data re-use

accommodates licensing

features access controls

allows for long term storage

osf.io

Accessible

Make the data and analysis accessible by

using formats that can be accessed with open sources software

documenting the formats of your files and how (i.e. with which software) they can be accessed

Interoperable

Make the data and analysis interoperable by

using / converting files to formats that are persistent, open source and software agnostic (e.g. .tsv, .csv, .txt)

Reusable

Make the data and analysis reusable by

providing detailed **documentation** of the contents of your files / data tables in a separate text file, which is often referred to as a **codebook**

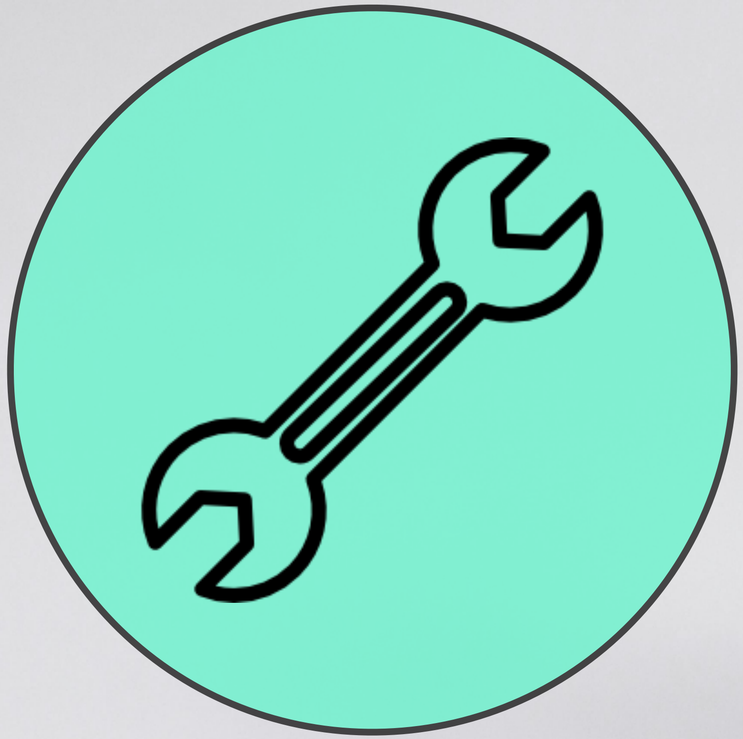
including a **description** of variables / functions and a **key to the labels** or abbreviations that are used to represent them

Specify an appropriate **licence** to inform others about the terms of reuse

GIT & GITHUB

version control system that takes “**snapshots**” of files, i.e. keeps track of incremental changes in files.

A **cloud-based hosting service** to manage Git where you can store projects & collaborate with others.



file/exercises /
03_exercise_github.pdf

Exercise 1: Make your first repo

Key ideas

repositories are projects

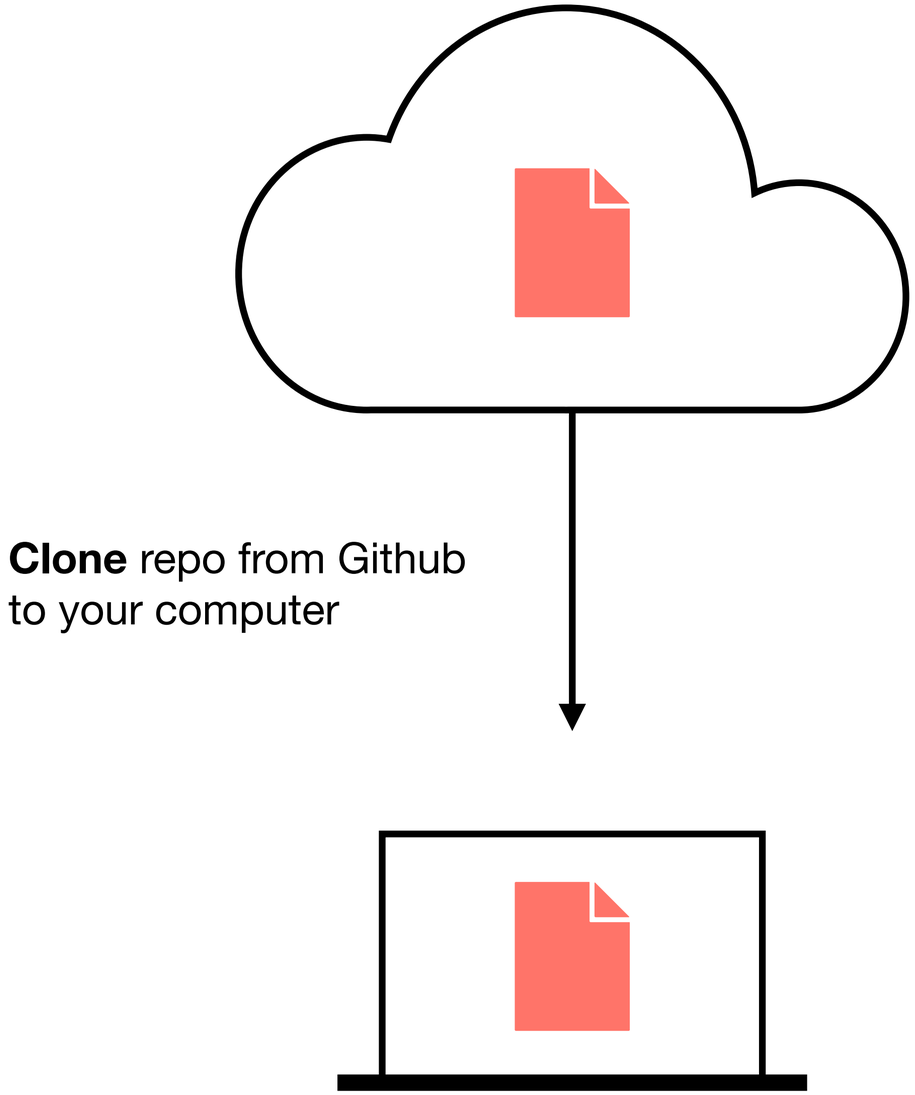
we **commit** changes to the repository

README files have special status on Github

Tips

think carefully about names

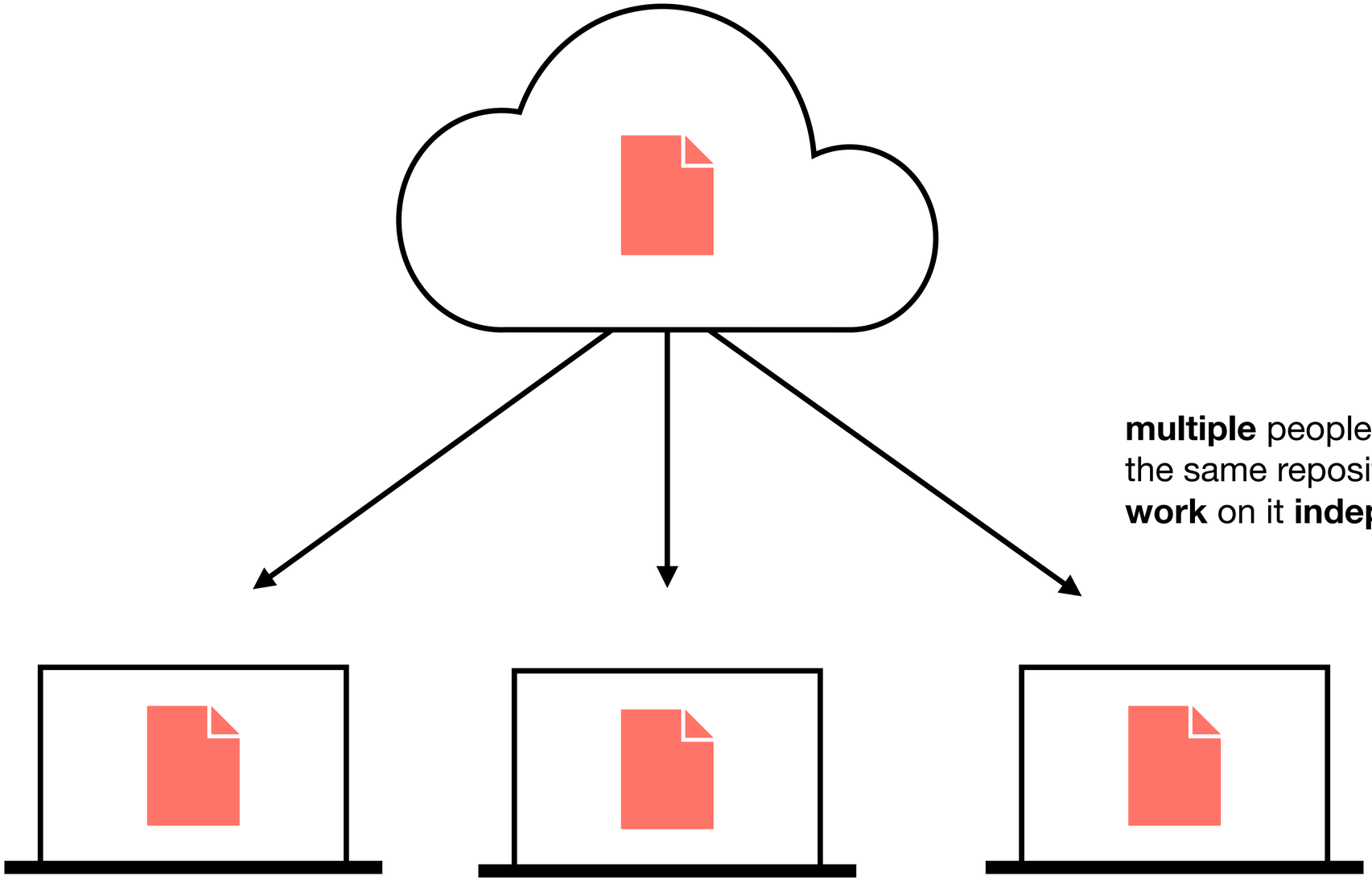
avoid spaces and uncommon characters



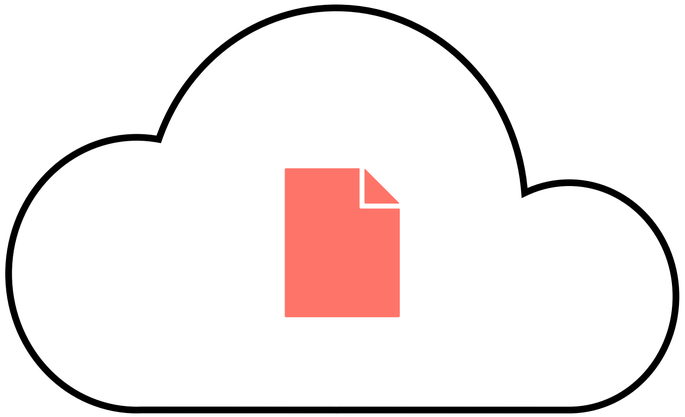
Your **remote repository** on Github

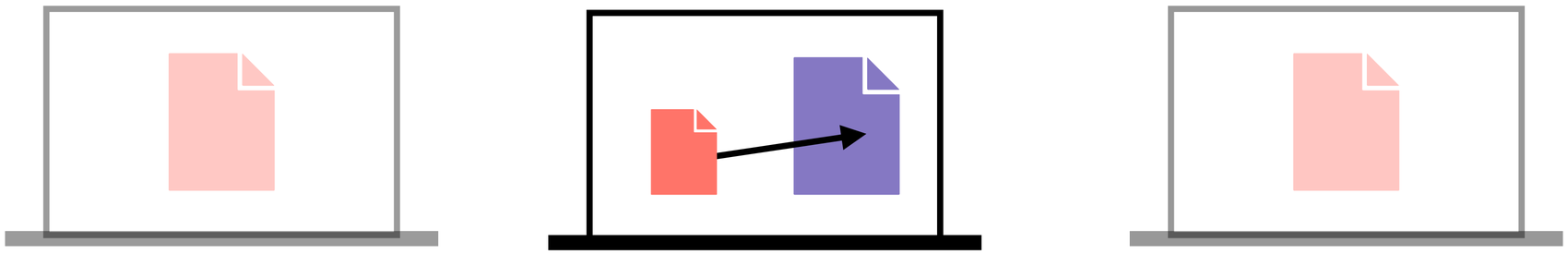
Clone repo from Github
to your computer

Your **local** copy on your machine

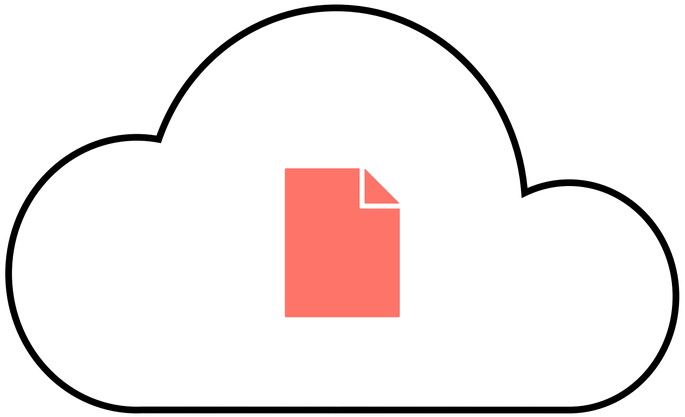


multiple people can clone the same repository and **work** on it **independently**

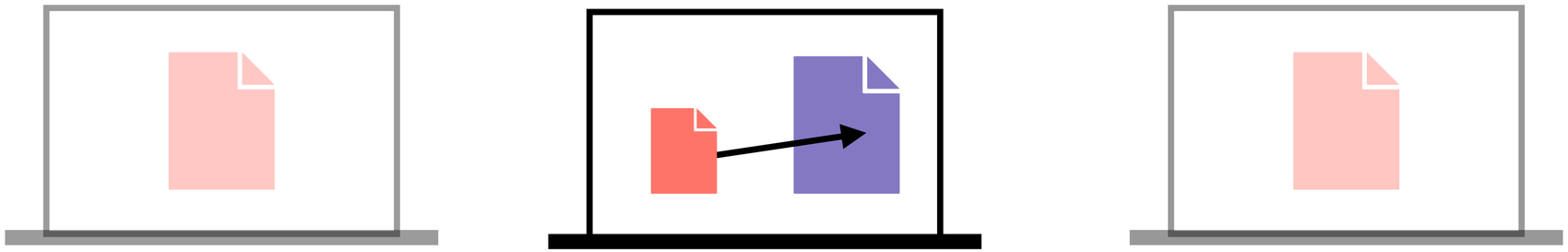




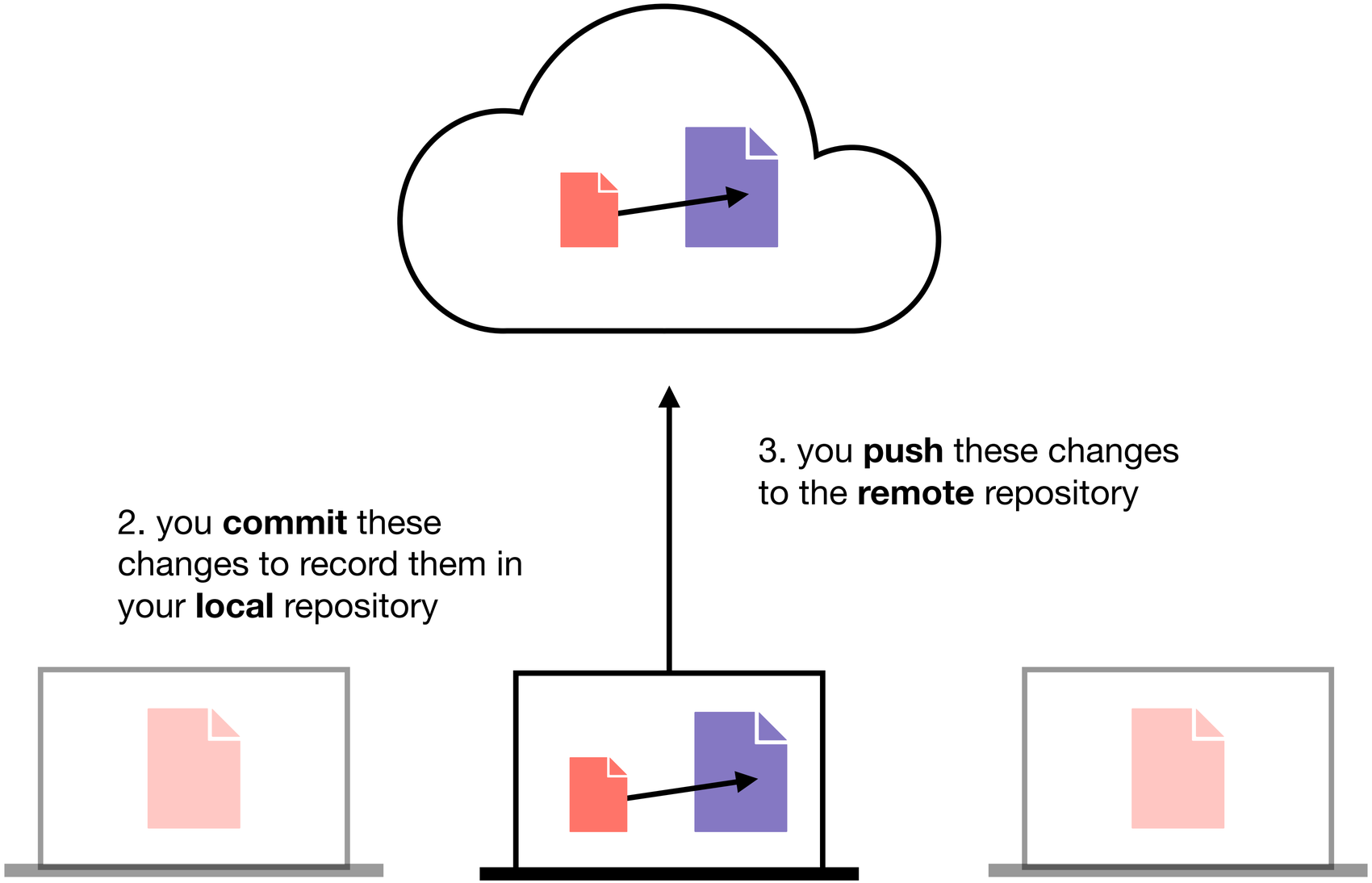
1. you **edit** the document
in your workspace

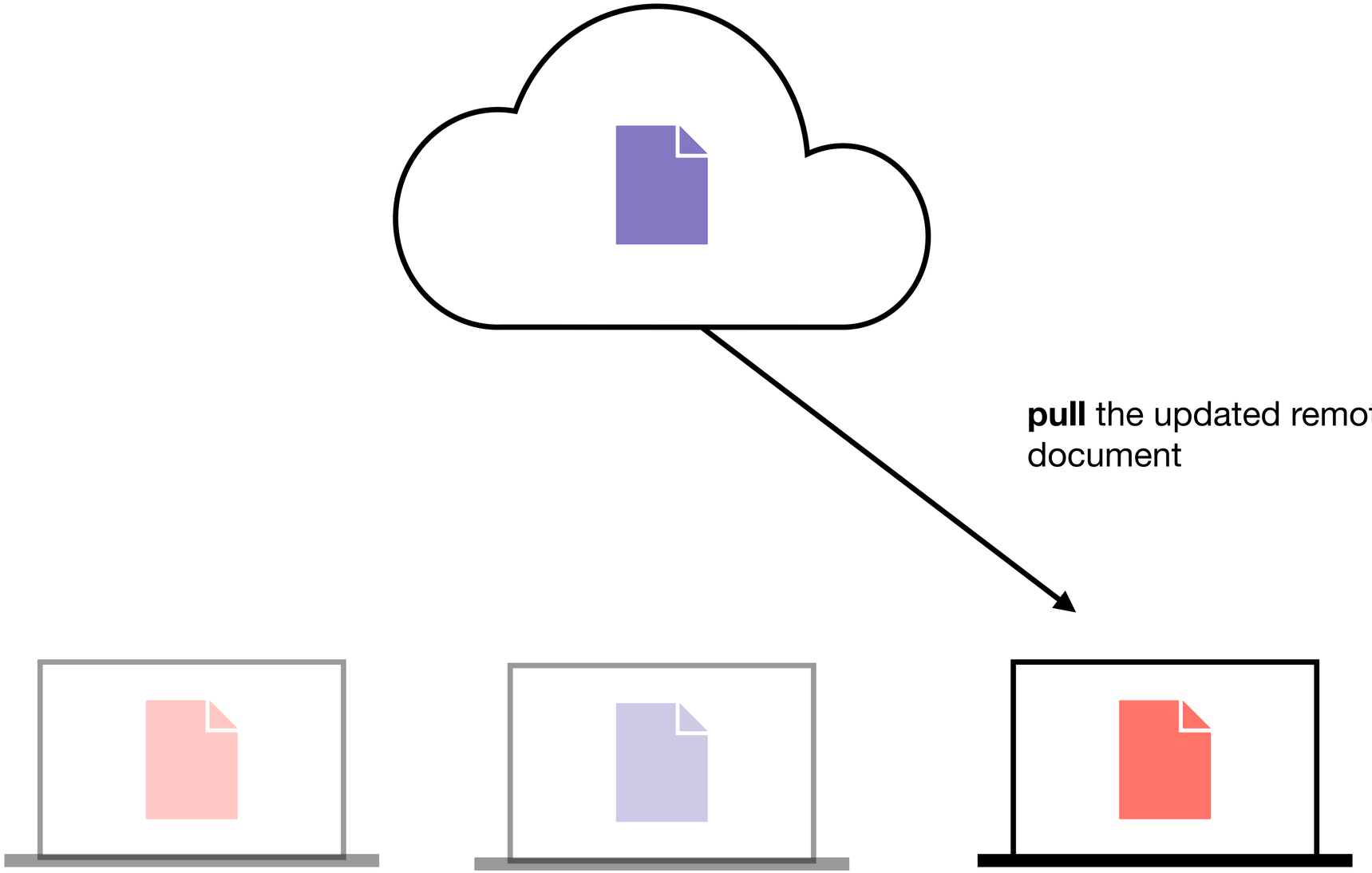


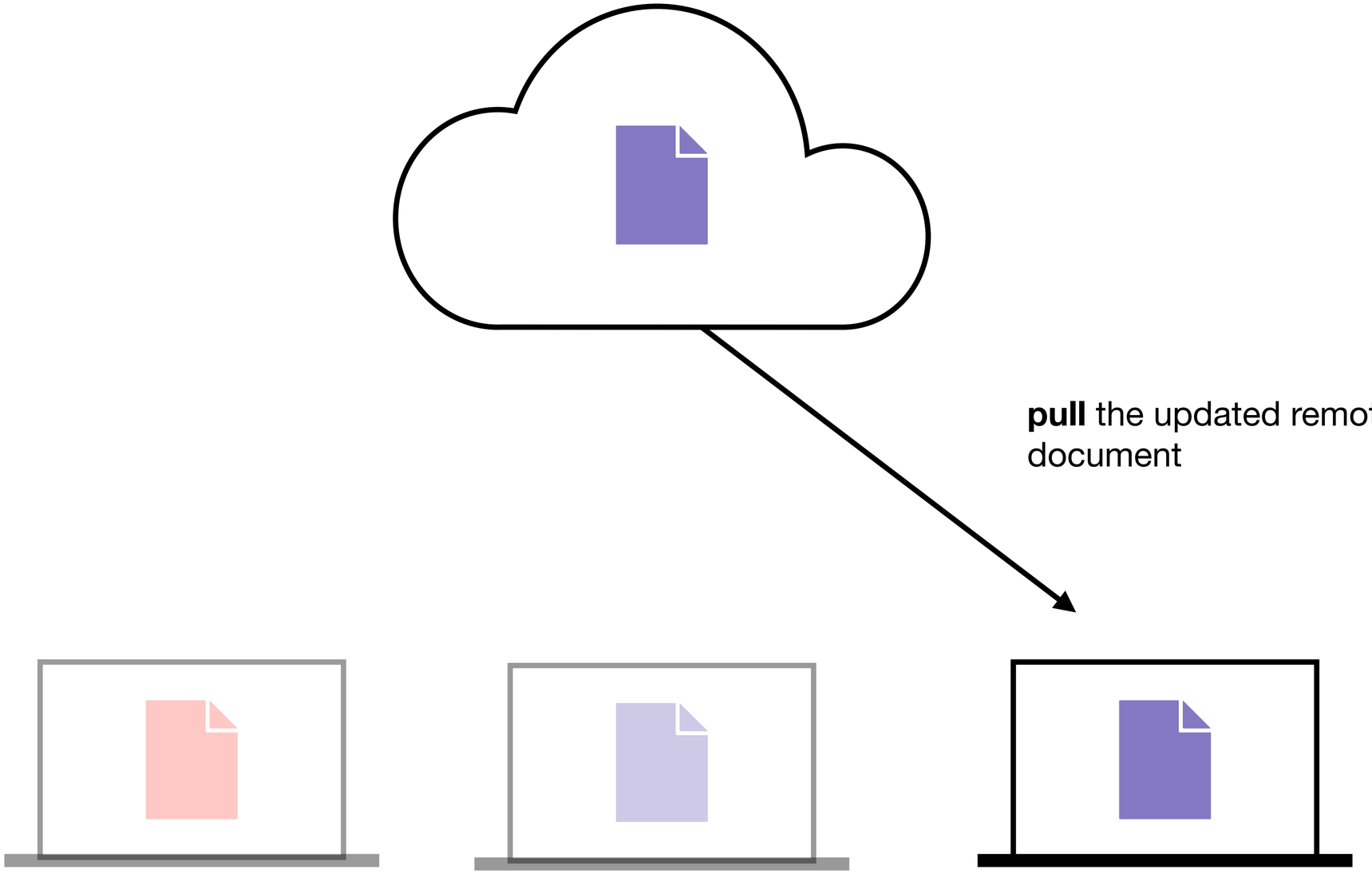
2. you **commit** these changes to record them in your **local** repository

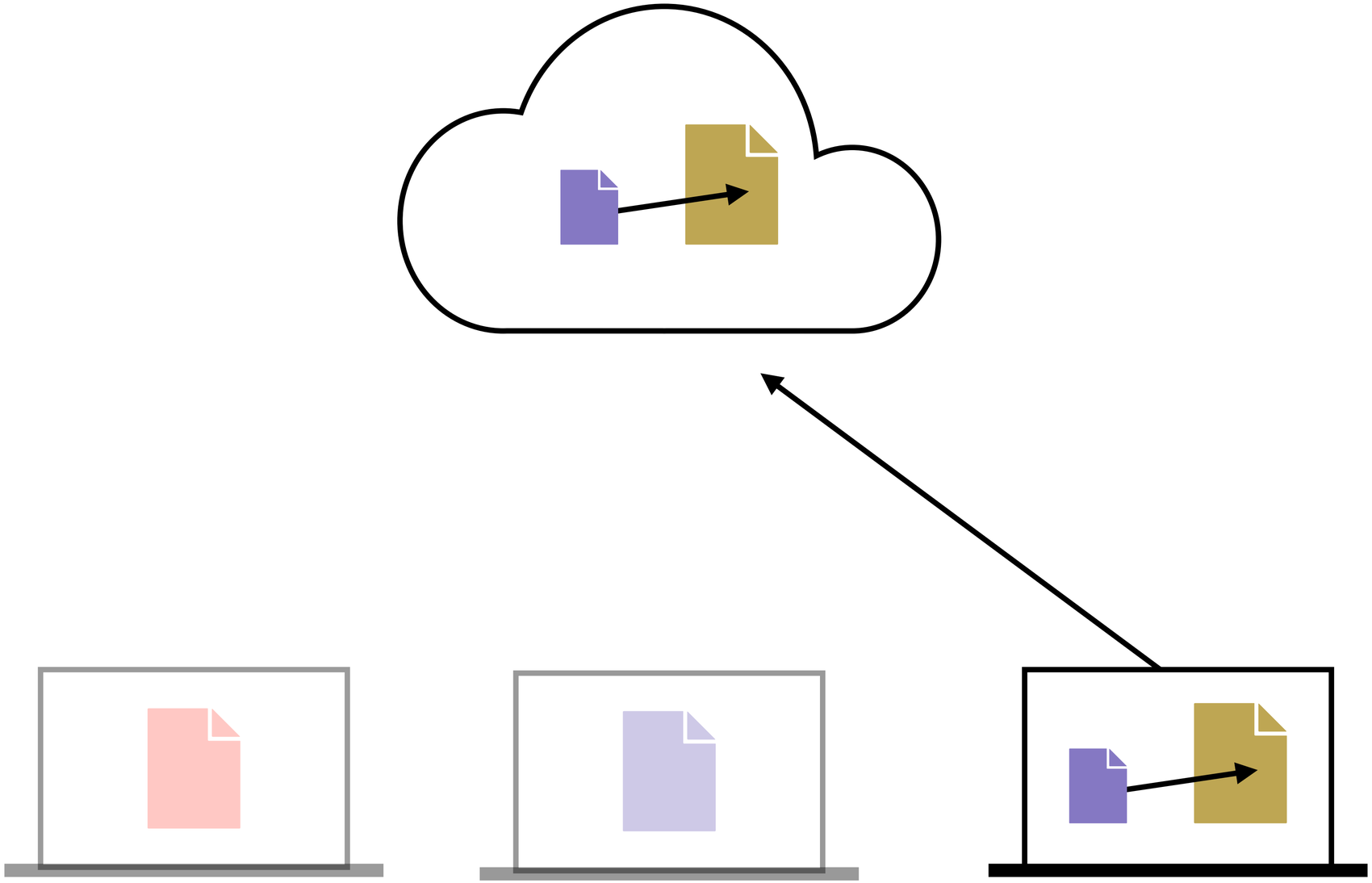


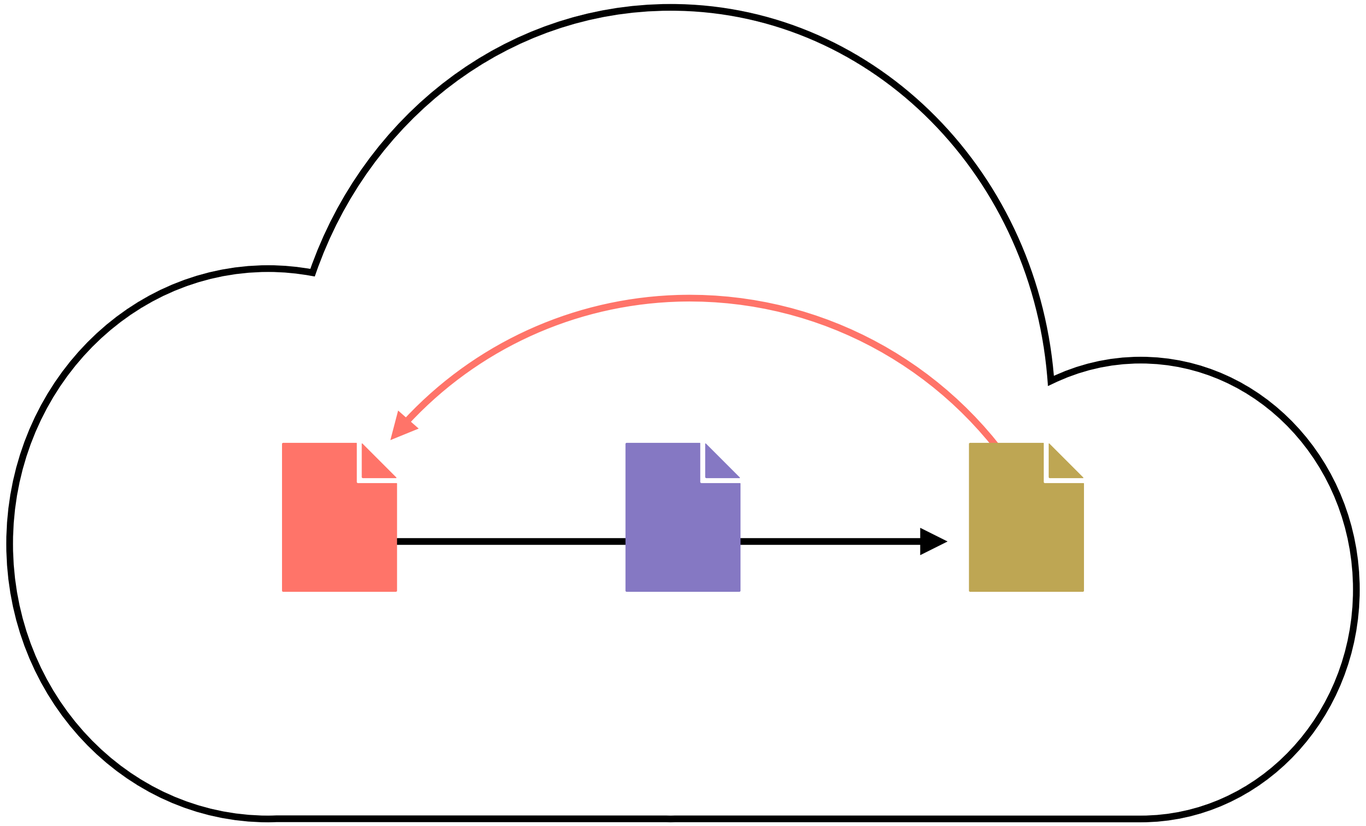
1. you **edit** the document in your workspace


2. you **commit** these changes to record them in your **local** repository
3. you **push** these changes to the **remote** repository
1. you **edit** the document in your workspace

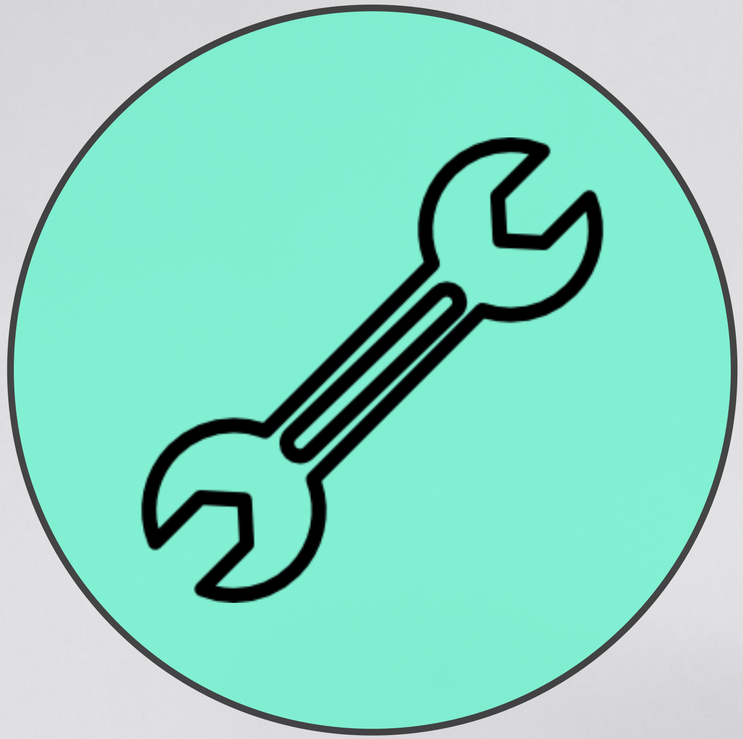

pull the updated remote
document


pull the updated remote
document


edit, commit, push

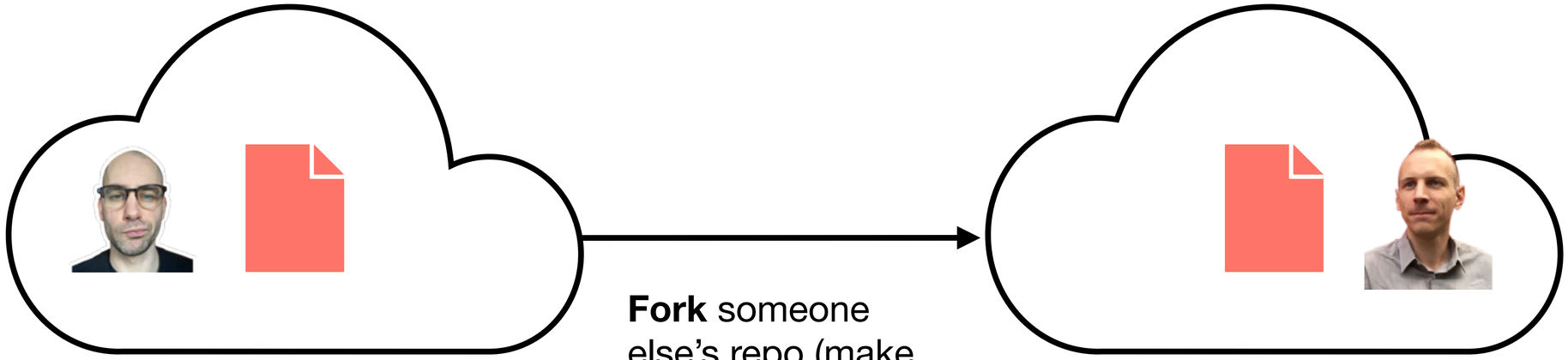


you can **revert** any changes to the remote repository to an earlier state.

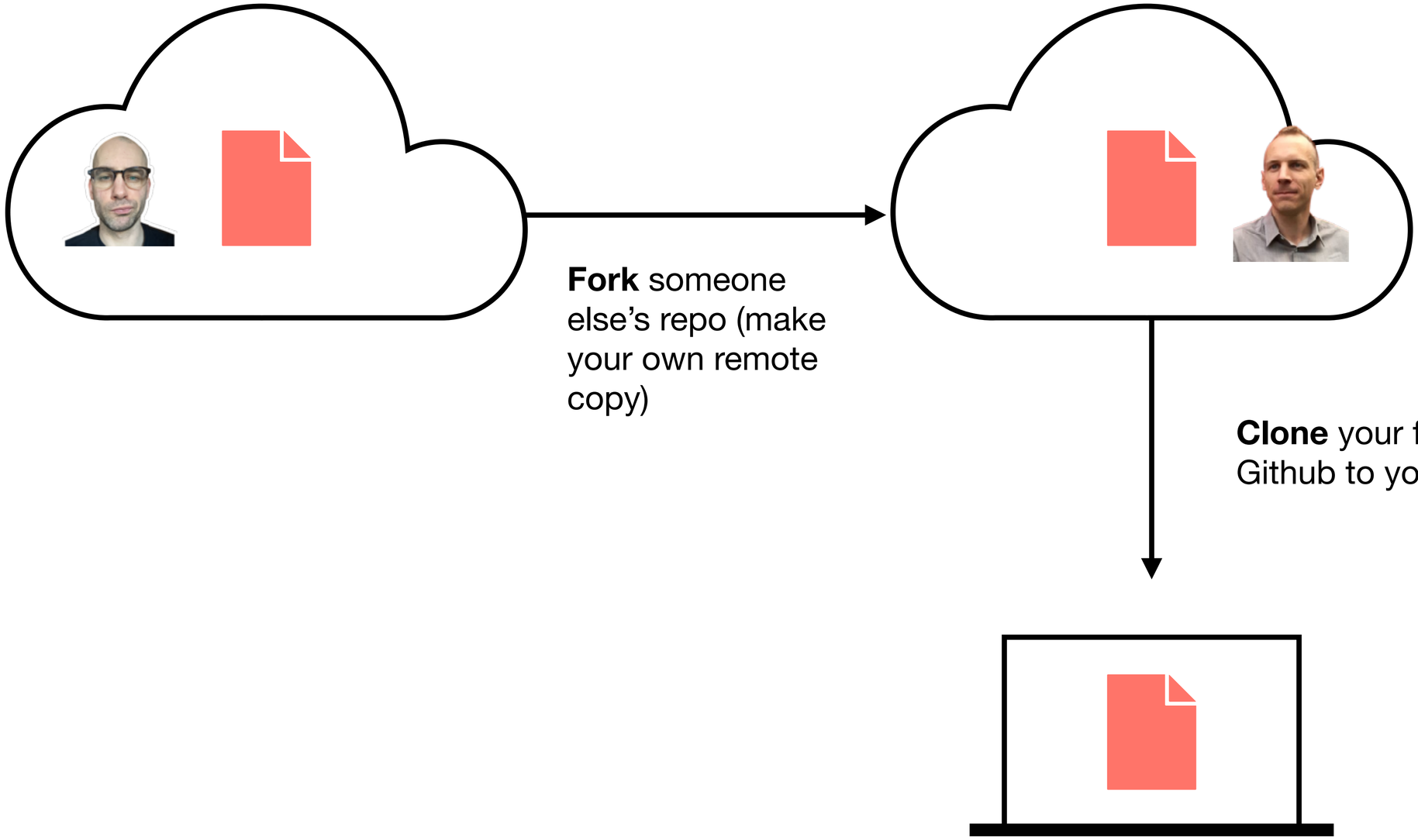


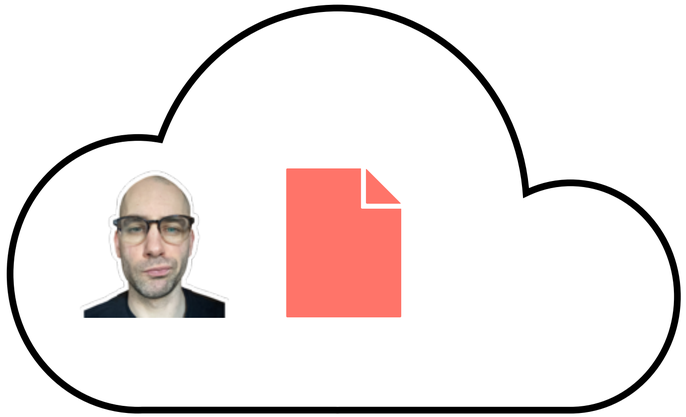
file/exercises /
03_exercise_github.pdf

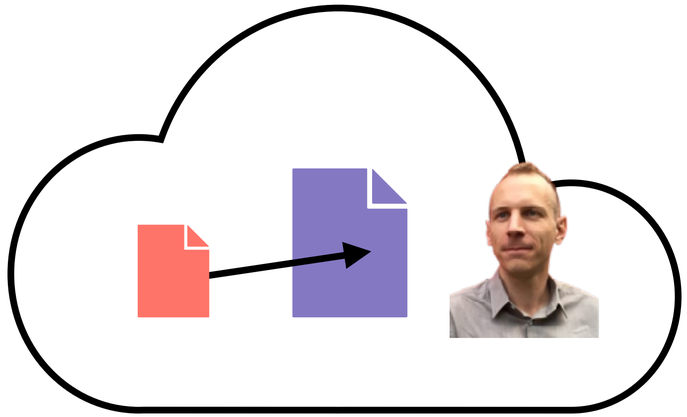
Exercise 2: Clone a repo

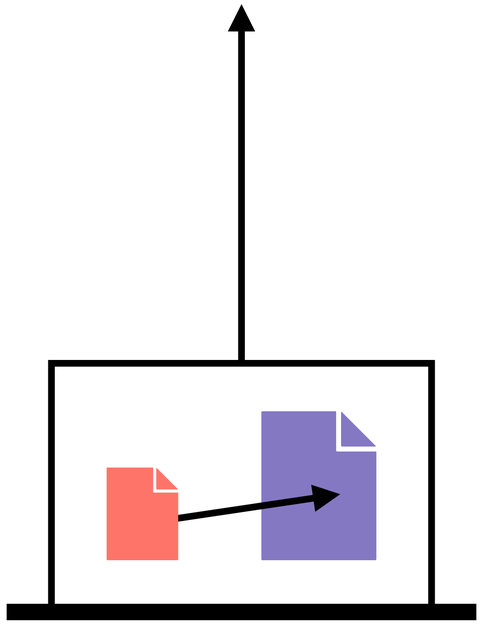


Fork someone else's repo (make your own remote copy)

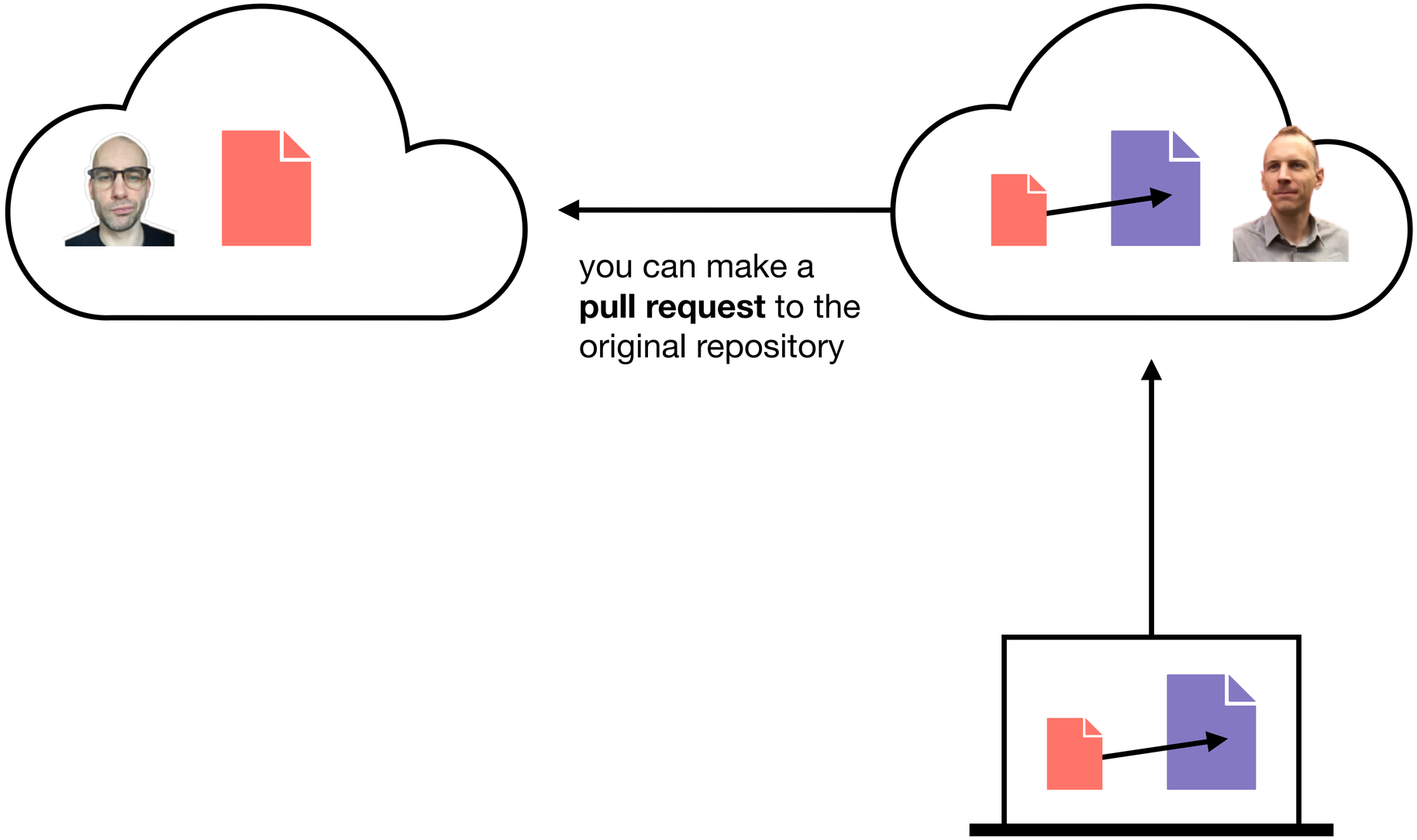

Fork someone else's repo (make your own remote copy)
Clone your fork from Github to your computer

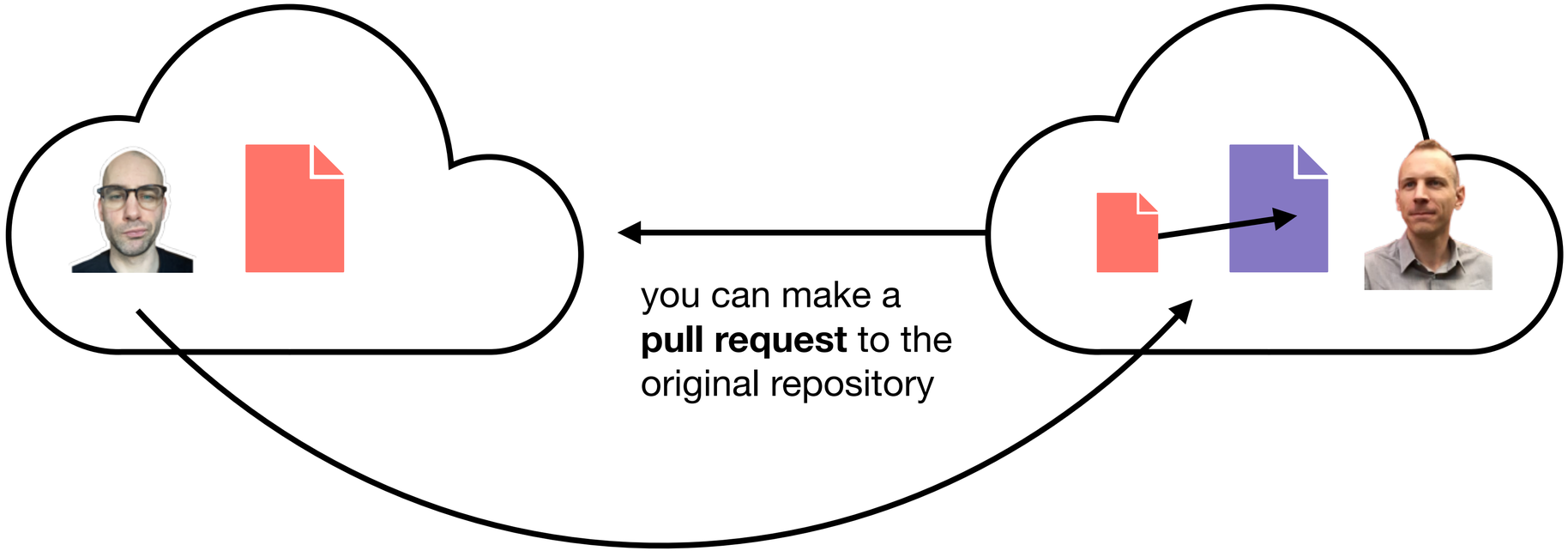






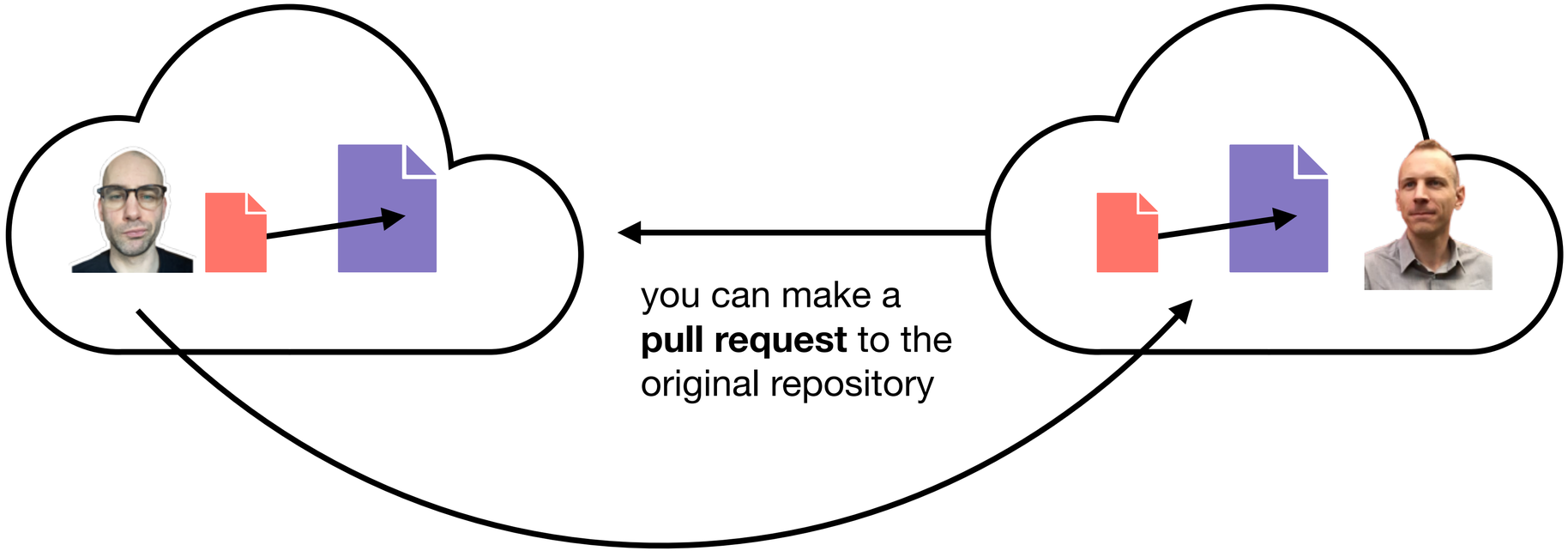
edit, commit, push


you can make a
pull request to the
original repository
edit, commit, push



original repo owner reviews the PQ
and either

rejects the pull request or



original repo owner reviews the PQ
and either

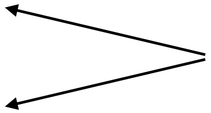
rejects the pull request or

accepts the pull request, integrating
the changes into the original repo

REPO STRUCTURE

RAW DATA

DERIVED DATA



Always separate raw data from derived data to make sure you do not overwrite the original data

SCRIPTS

PLOTS