Practical Steps Towards Reproducible Research

Heidi Seibold / @HeidiBaya

Trainer + Consultant for Open and Reproducible Data Science

WORKSHOP

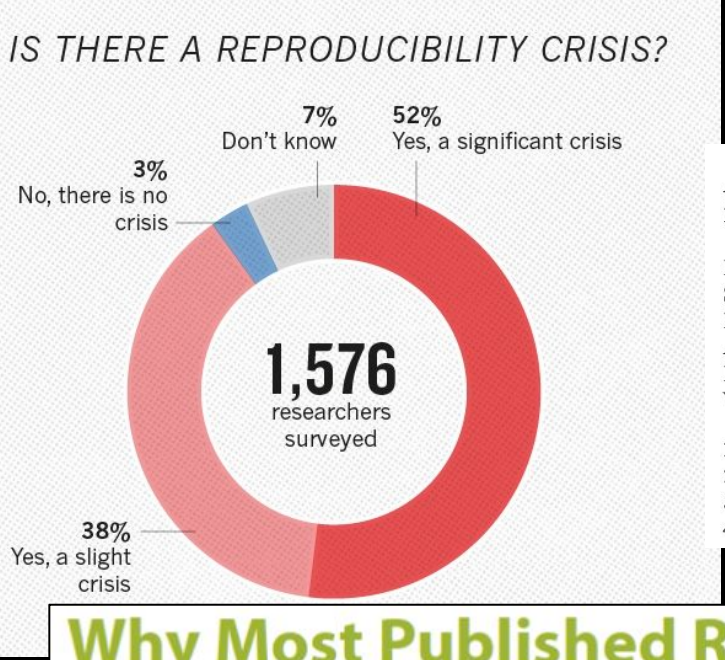
LOVE YOUR DATA? MAKE IT REPRODUCIBLE!

Tuesday, 14.02.2023 · 09:30 am - 01:00 pm online


Please ask questions in Miro

Why are you here?

→ Breakout room (5 minutes)



A computational reproducibility study featuring longitudinal data analyses

Heidi Seibold[*,1,2,3,4], Severin Czerny[1], Siona Decke[1], R
Steffen Fohr[1], Nico Hahn[1], Rabea Hartmann[1], Christo
Dario Lepke[1], Verena Loidl[1], Maximilian Mandl[1], Sara
Alexander Piehler[1], Elio Rojas[1], Stefanie Schmid[1], Han
Lennart Schneider[1], Xiao-Yin To[1], Viet Tran[1], Antje V
Wagner[1], Maria Waize[1], Hannah Wecker[1], Rui Yang[1],

1 LMU Munich, Germany
2 University of Bielefeld, Germany
3 Helmholtz Zentrum München, Germany
4 LMU Open Science Center

Why Most Published Research Findings Are False

John P. A. Ioannidis

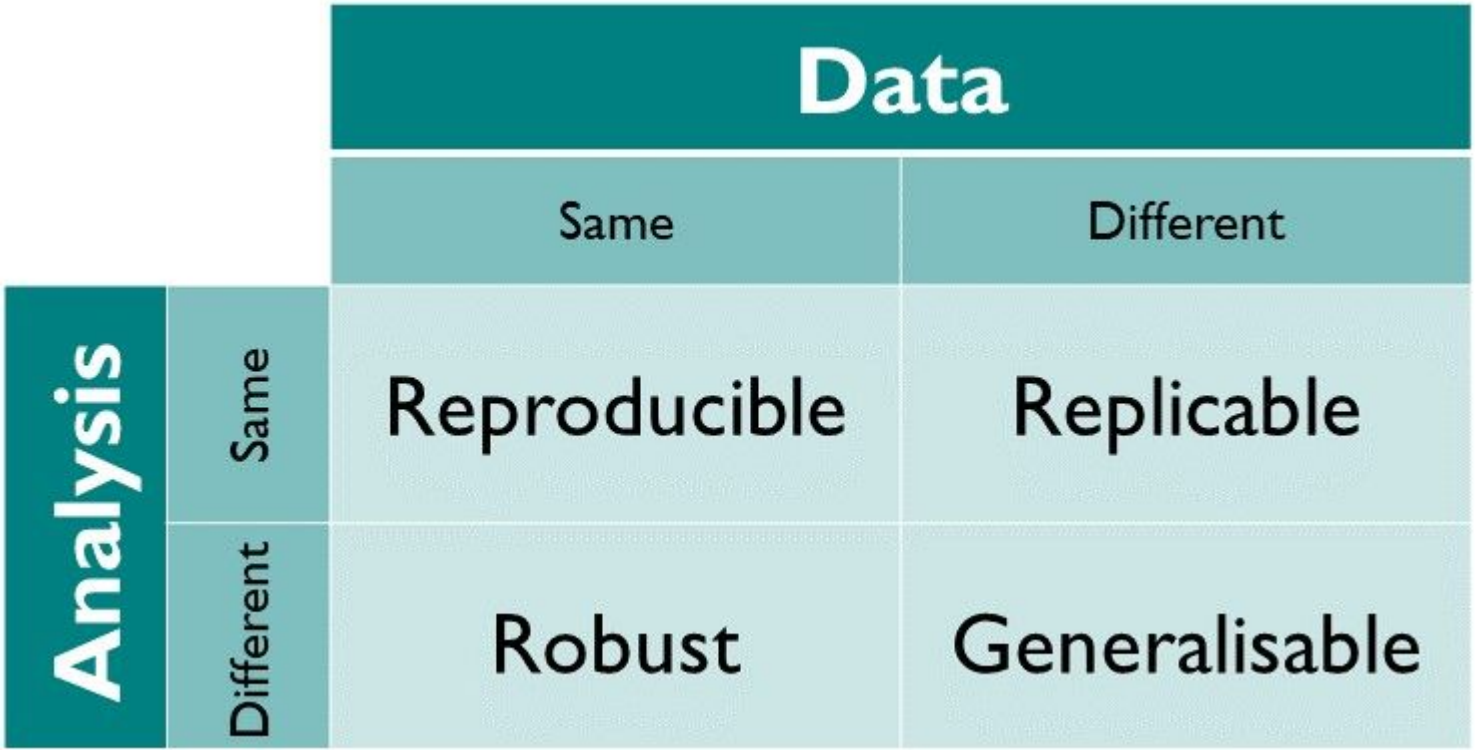
DOI: 10.1371/journal.pmed.0020124

Is this what we want?

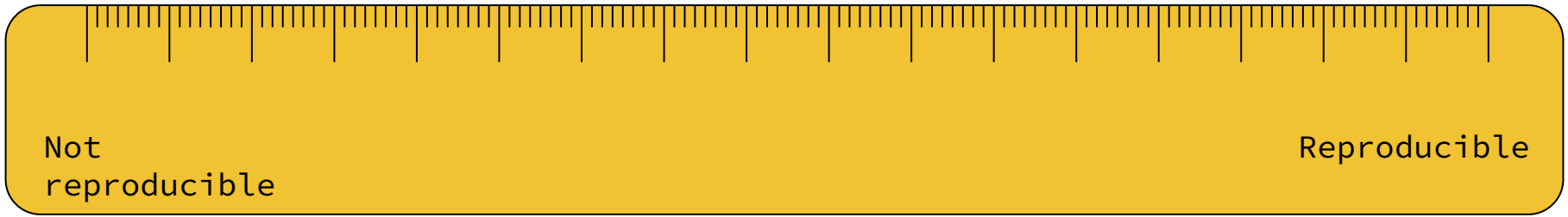
How to become a science champion?

What makes a science champion?

→ Answer in Miro



From [The Turing Way]


Not
reproducible
Reproducible

Practical steps towards computational reproducibility



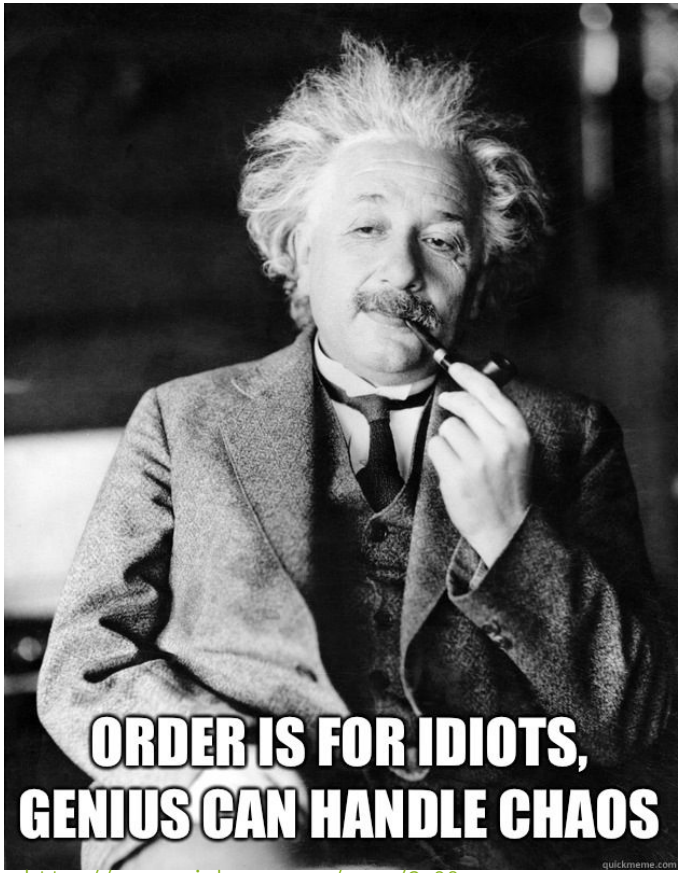



Practical steps towards reproducible research

- Get organized!
- Use code
- Use version control
- Make your work available online

Get organized!

Good organisation



http://www.quickmeme.com/meme/3r98zx

Let's not pretend: we're not geniuses ;P

Good organisation
… starts simple

- No sending emails with files
- Nice file organisation
- Good naming

Dear colleagues,

attached you find the first public version of the [redacted] protocol. Please have a look and do comment. We can also meet to aggregate our reviews.

1 attachment: StudyProposa[redacted]_Validation_V1_250918docx.docx

```
├── Makefile
├── references_topic.bib
├── manuscript_topic.Rnw
├── abstract.tex
├── introduction.tex
├── methods.tex
├── application.Rnw
├── simulation.Rnw
├── discussion.tex
├── journal_correspondence
│   ├── journal2
│   │   ├── response1.txt
│   │   └── p2p_1.tex
│   └── journal1
│       └── response.txt
└── analysis
    ├── Makefile
    ├── try_stuff.Rmd
    ├── simulation.Rnw
    ├── simulation.R
    ├── analysis_study1.R
    ├── analysis_study2.R
    ├── data
    │   ├── clean_data.R
    │   ├── data_study2.csv
    │   ├── data_study1.csv
    │   └── simulation_results.RData
    ├── data_original
    │   ├── data_study1_original.csv
    │   └── data_study2_original.csv
    └── basis
        ├── functions_simulation.R
        └── functions_analyses.R
```

Naming

NO

- Myabstract.docx
- Joe’s Filenames Use Spaces and Punctuation.xlsx
- figure 1.png
- fig 2.png
- JW7d^(2sl@deletethisandyourcareerisoverWx2*.txt

YES

- 2014-06-08_abstract-for-sla.docx
- Joes-filenames-are-getting-better.xlsx
- Fig01_scatterplot-talk-length-vs-interest.png
- Fig02_histogram-talk-attendance.png
- 1986-01-28_raw-data-from-challenger-o-rings.txt

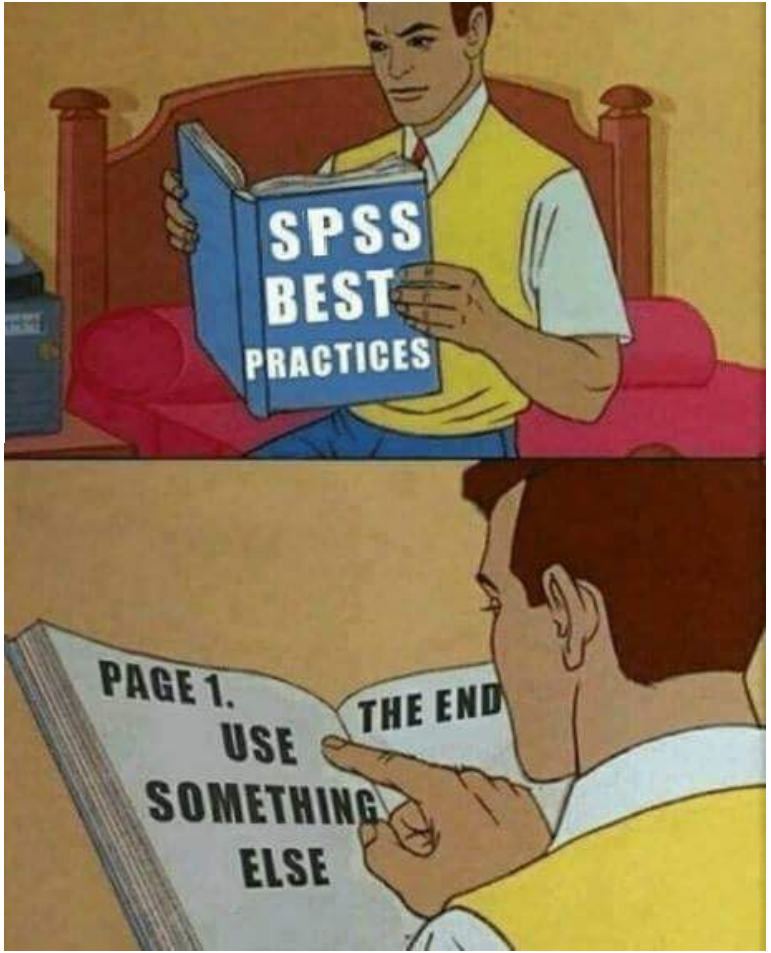
See slides by Jenny Brian

Use Code

https://www.nature.com/articles/d41586-021-02211-4

Autocorrect errors in Excel still creating genomics headache

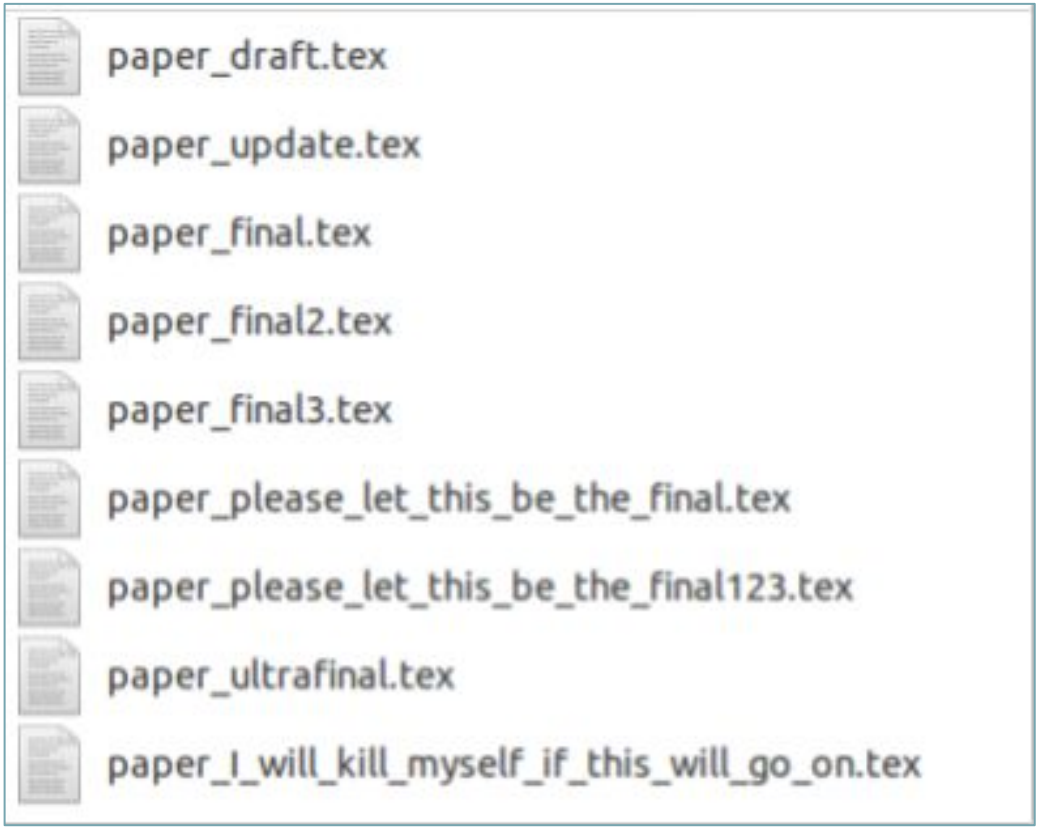
Despite geneticists being warned about spreadsheet problems, 30% of published papers contain mangled gene names in supplementary data.



https://devrant.com/rants/686860/spss-101

Use Version Control

Version Control in the olden days



Real Version Control (Including Backup)

Commit fe6b5538 authored 10 months ago by **HeidiSeibold** Browse files Options

fix typos

parent e8a0da8a master

No related merge requests found

Showing 1 changed file with 3 additions and 3 deletions Hide whitespace changes Inline Side-by-side

2018_swisscore.tex View file @ fe6b5538

```
...  ...  @@ -56,8 +56,8 @@ In my research I develop algorithms that detect which patient characteristics
56   56   lead to a positive or negative reaction to a therapy. The algorithms can also
57   57   detect and predict which patients are likely to have side effects from a
58   58   therapy. It is important to acknowledge, that \textbf{personalised medicine does not
59        - neccessarily mean, that each patient receives a personalised treatment, but it
60        - can also mean that we find that treatments work similarly accross patients and
     59   + necessarily mean, that each patient receives a personalised treatment, but it
     60   + can also mean that we find that treatments work similarly across patients and
```

Prerequisites for Version Control

- **Text**
  - Documents/Papers: LaTeX, Markdown
  - Analyses: R, Python

- **Willingness to learn something new**: Git

Make your work available online






FAIR Data

Data and supplementary materials have sufficiently rich metadata and a unique and persistent identifier.

FINDABLE

Metadata and data are understandable to humans and machines. Data is deposited in a trusted repository.

ACCESSIBLE

Metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation.

INTEROPERABLE

Data and collections have a clear usage licenses and provide accurate information on provenance.

REUSABLE

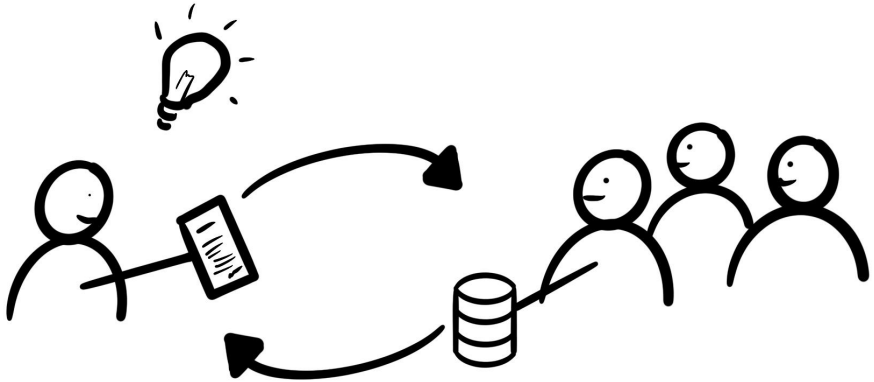
https://libereurope.eu/article/fairdataconsultation/

Open Code

```
# Steps

s1 <- "Use scripts (no clicking!)"

s2 <- "Publish code online"


# Pro tip

s3 <- "Use version control"
```

Open Materials

Open Science Framework
A scholarly commons to connect the entire research cycle



zenodo

GitLab

Break time

Practical steps towards reproducible research

- Get organized!
- Use code
- Use version control
- Make your work available online

How hard can it be?

1. Draw an animal on a piece of paper
2. Write instructions on how to draw what you drew (add them to Miro)
3. Go to breakout room → draw according to the instructions of the other person
4. Check your result: did you succeed?
5. If you want: upload the original and the reproduced drawing to the Miro
6. Come back to the main room

How hard can it be?

- What can we learn from that exercise?
- Is it easier or more difficult in your work?

Be a science champion + be selfish

Avoid disaster

Write papers more easily

Convince reviewers

Facilitate continuity of work

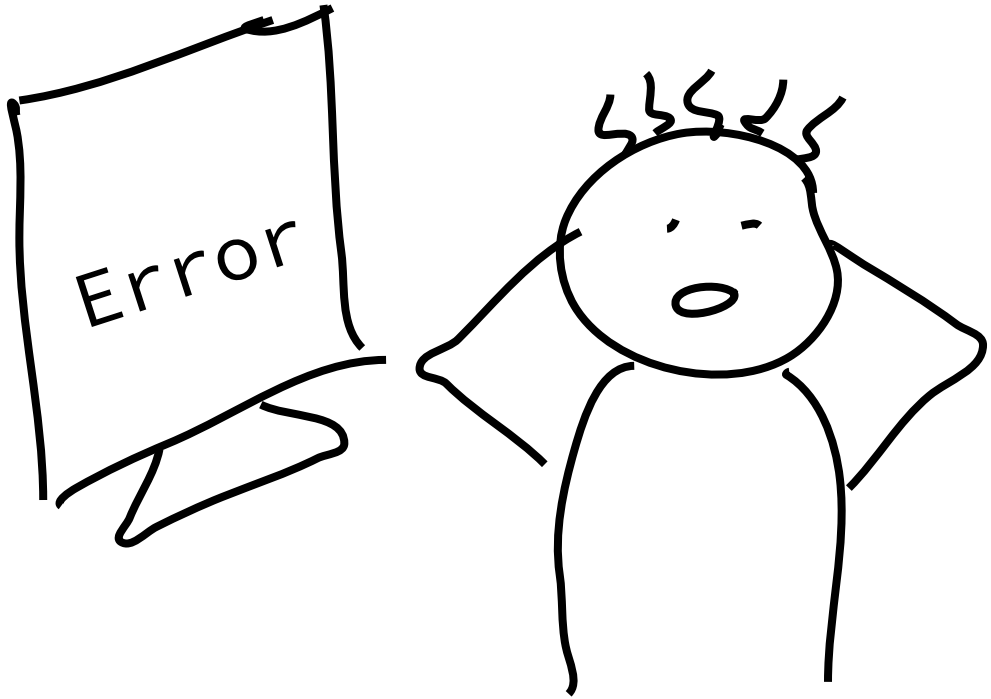
Build your reputation

Markowetz (2015). Five selfish reasons to work reproducibly. https://doi.org/10.1186/s13059-015-0850-7

What is stopping you / others?

→ Answer in Miro

Overwhelmed?

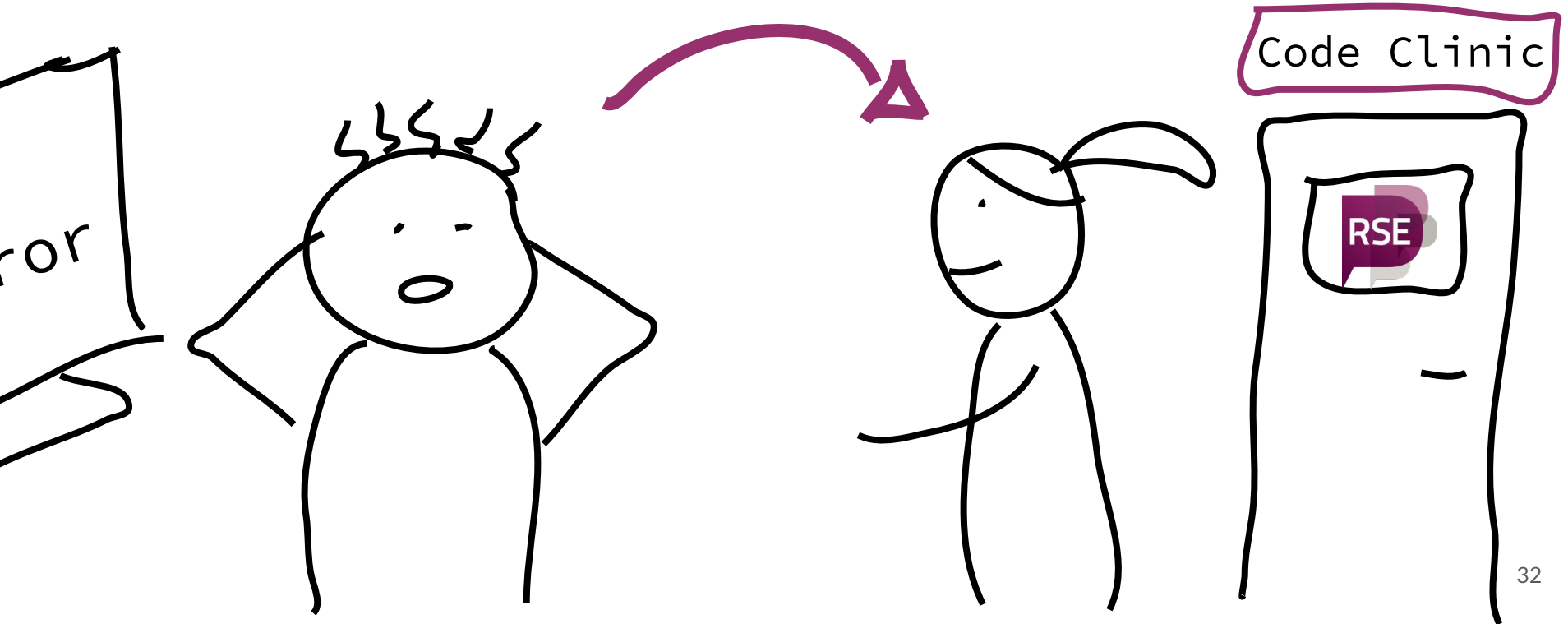


Reproducible Research

requires

Software Skills

Research Software Engineers can help



Help us help you

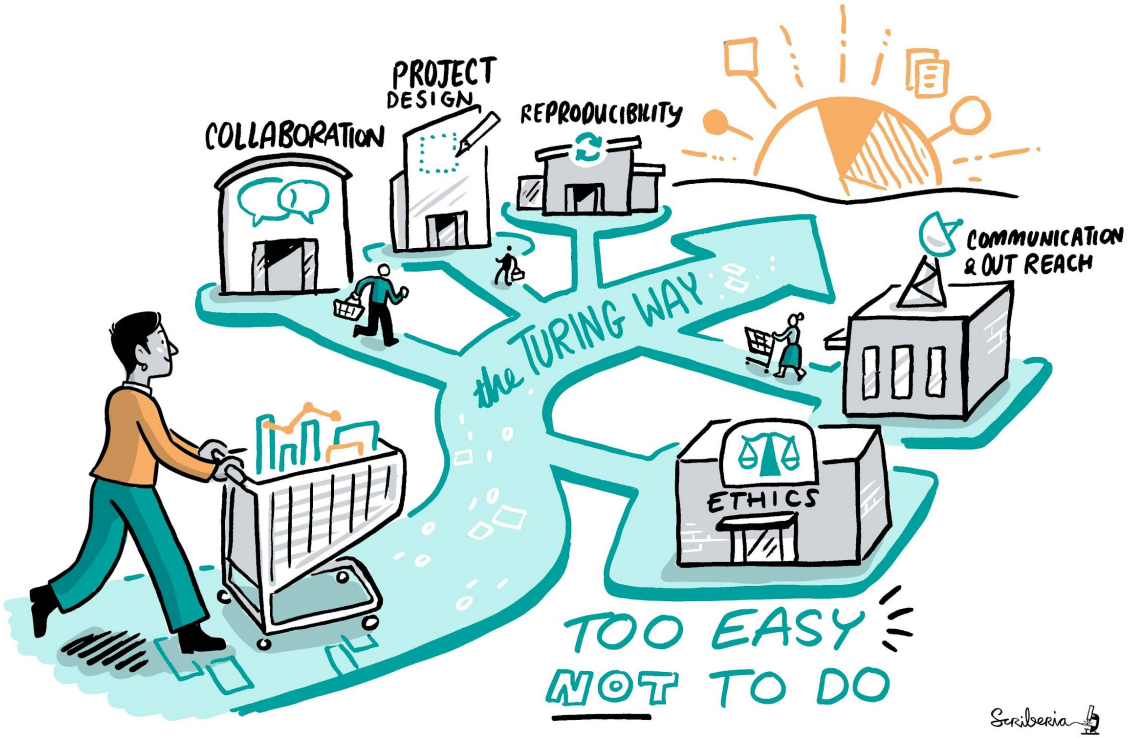
de RSE | GESELLSCHAFT FÜR FORSCHUNGSSOFTWARE

Summary time

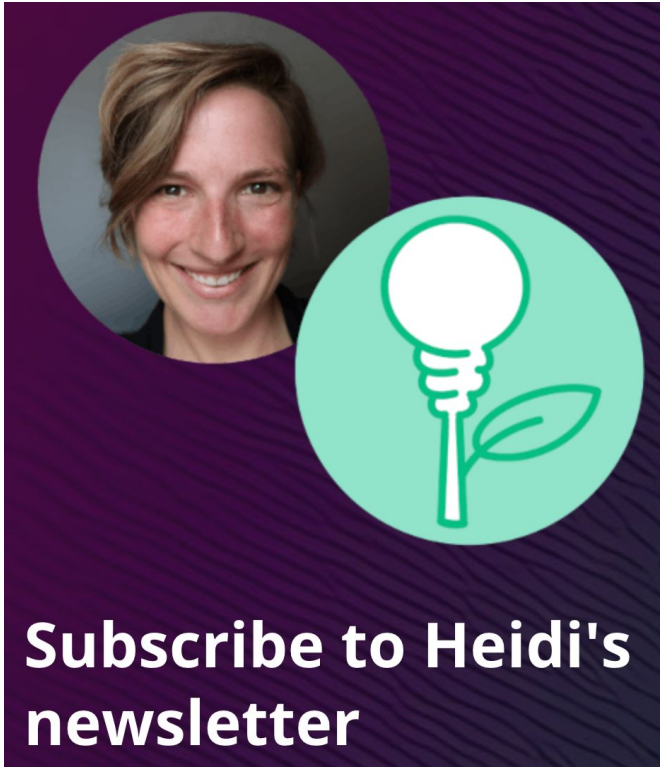
Sum up what you've learned,
write down your next step

(creativity welcome 😉)

Further reading



the-turing-way.netlify.app



https://heidiseibold.ck.page

ASK ME ANYTHING

AMA: open & reproducible data science

You send in questions and I will answer them as truthfully and helpfully as possible.

You can ask about anything revolving around open and reproducible data science: open data, open source, good coding practices, licenses, fears, social dilemmas, ...

https://heidiseibold.ck.page/amalovedata