

Open Science and Research Software Engineering

Workshop
Center for Advanced Internet Studies (CAIS)

Quirin Würschinger

September 21, 2023

Introduction

Workshop materials

GitHub repository <https://github.com/wuqui/opensciws>
slides https://wuqui.github.io/opensciws/opensciws_slides.html
website version https://wuqui.github.io/opensciws/opensciws_website.html

> **whoami**

Quirin Würschinger
q.wuerschinger@lmu.de
Wissenschaftlicher Mitarbeiter and PostDoc in (computational) linguistics
LMU Munich

Current work

- **research**
 - lexical innovation on the web and in social networks
 - variation and change in language use and social polarization in social networks



- using Large Language Models (LLMs) like ChatGPT for research in linguistics and social science.
- **teaching:** corpus linguistics and research methodology

Promoting Open Science in (computational) linguistics at LMU

- teaching and applying reproducible corpuslinguistic methods
 - creating and sharing corpora among researchers and students
-

Topics

- **Open Science:** What is it and why does it make sense?
 - **Project setup:** Learn to efficiently organize your project's files and folders.
 - **Data, code, and document management:** Understand how to effectively manage and maintain your data, code, and documents.
 - **Utilizing different data types:** Discover how to work with various data types, such as interviews, web, and social media data.
 - **Publication process:** Learn about options for publishing your data, code, and documents.
-

Time table

Topic	Start	End
Intro	09:00	09:30
Open Science principles	09:30	10:30
—	10:30	10:50

Topic	Start	End
version control	10:50	11:10
project structure	11:10	12:00
data	12:00	12:30
—	12:30	13:30
code	13:30	14:00
methods	14:00	14:30
authoring	14:30	15:15
—	15:15	15:30
publishing	15:30	16:00
open issues and recap	16:00	16:30

Addressing different backgrounds and goals

Backgrounds and interests

CAIS: Forschung zu Digitalisierung und Digitale Gesellschaft

Fachrichtungen

- Politikwissenschaft
- Erziehungswissenschaft
- Kommunikationswissenschaft
- Soziologie
- ...

Daten und Methoden

- qualitativen Interviews
 - Textanalysen
 - quantitativen Befragungen
 - experimentellen Designs
 - Online-Plattformen/Social Media
 - ...
-

Main interests

- reproducible workflows
 - managing files and folders
 - plain text authoring
 - programming with Python and R
 - methods
 - quantitative approaches
 - text analysis
 - questionnaires
 - publishing
 - papers
 - code and data
-

Who are you?

1. name
 2. place and position
 3. your research interest in about 3 sentences for someone outside your field
-

Open Science principles

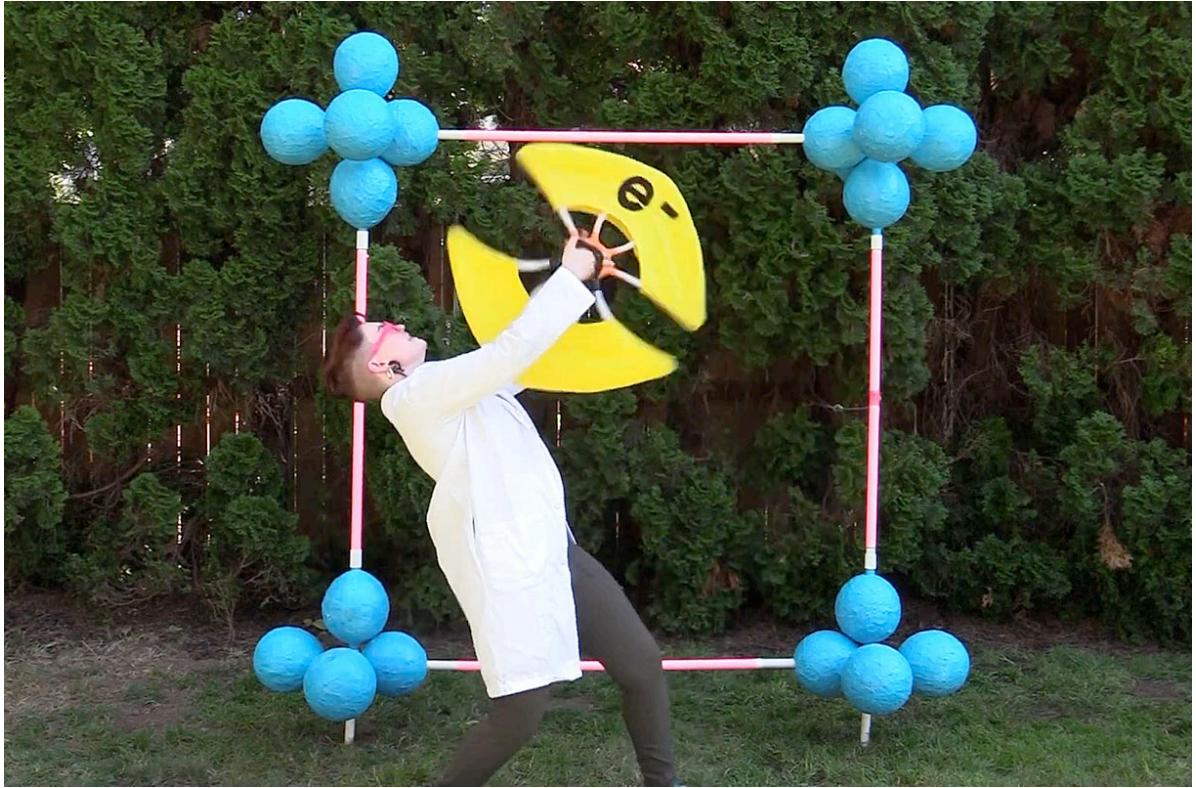
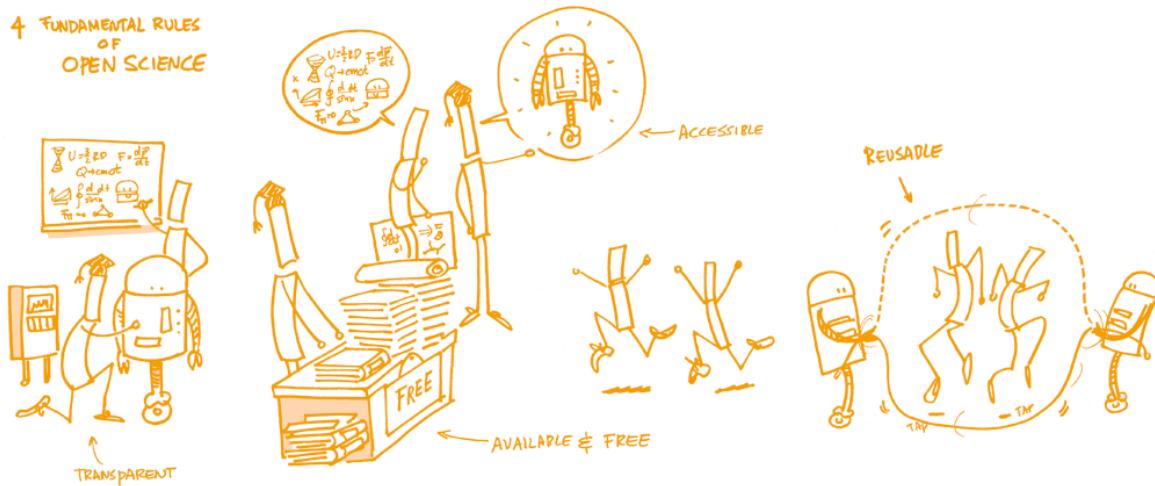


Figure 1: [Dance your PhD](#)

What is Open Science



Why should we do Open Science?



Is this what we want?

Scientists: Mostly Hackers

“The case against science is straight-forward: much of the scientific literature, perhaps half, may simply be untrue. Afflicted by studies with small sample sizes, tiny effects, invalid exploratory analyses, and flagrant conflicts of interest, together with an obsession for pursuing fashionable trends of dubious importance, **science has taken a turn towards darkness.**”

Richard Horton



Richard Horton, United Kingdom
Editor-in-Chief
The Lancet

Figure 2: Richard McElreath: *Science as Amateur Software Development*

Hack Your Way To Scientific Glory

You're a social scientist with a hunch: **The U.S. economy is affected by whether Republicans or Democrats are in office.** Try to show that a connection exists, using real data going back to 1948. For your results to be publishable in an academic journal, you'll need to prove that they are "statistically significant" by achieving a low enough p-value.

P-hacking

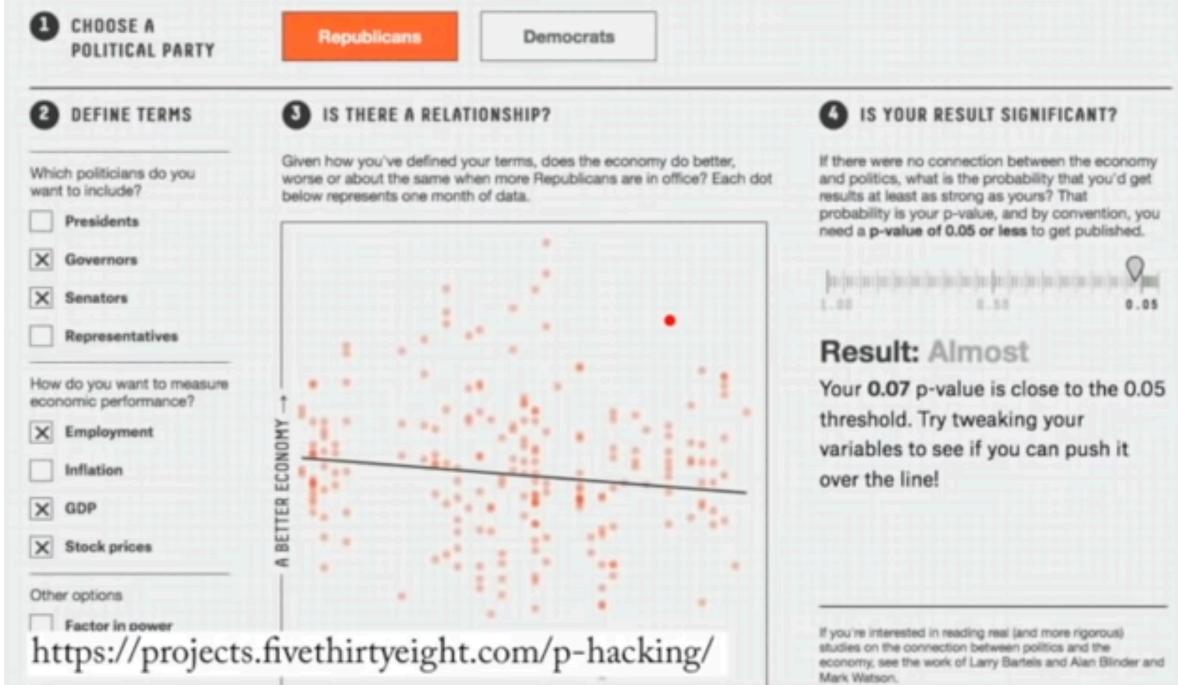
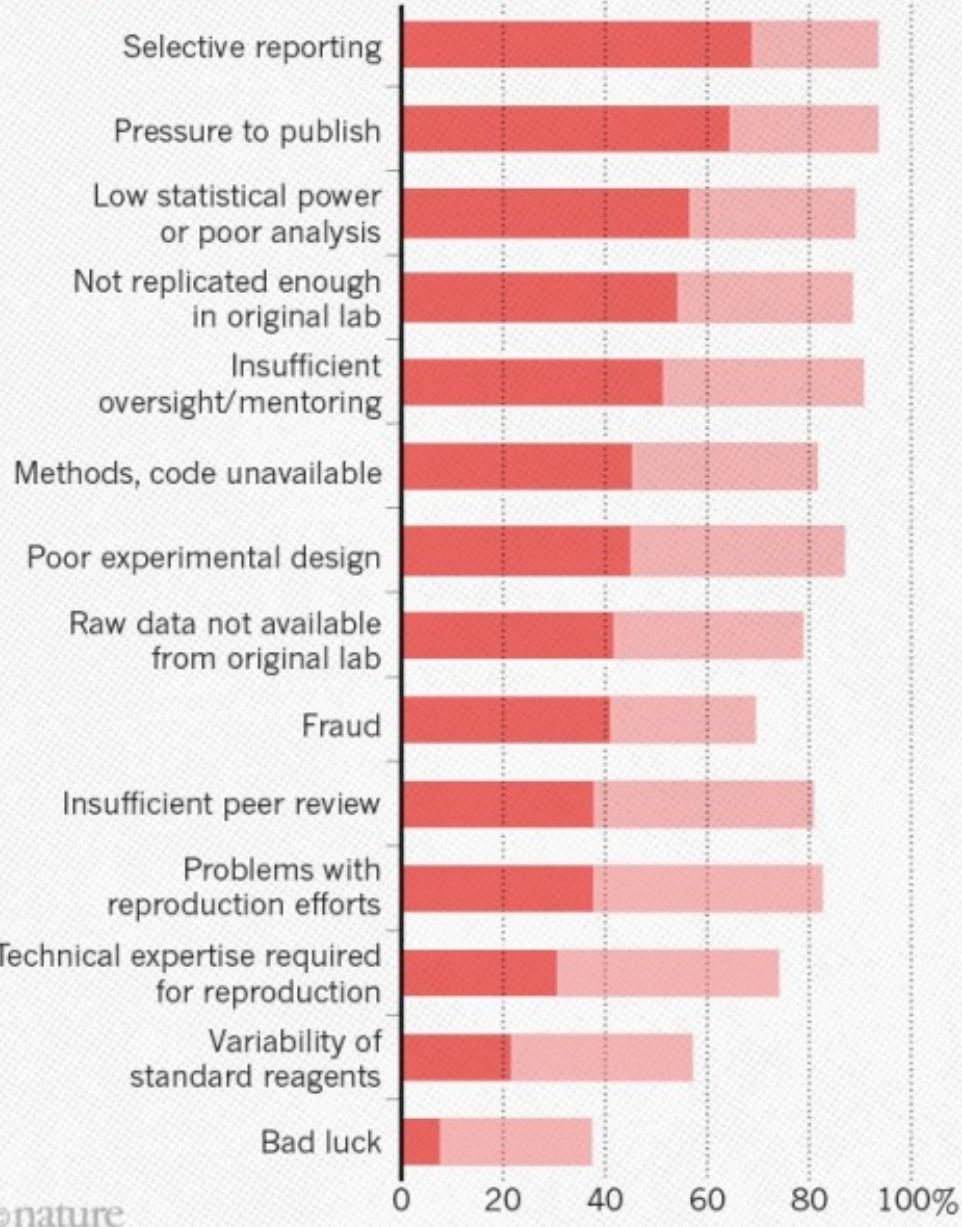


Figure 3: [source](#)

WHAT FACTORS CONTRIBUTE TO IRREPRODUCIBLE RESEARCH?

Many top-rated factors relate to intense competition and time pressure.

● Always/often contribute ● Sometimes contribute



©nature

Figure 4: [source](#)

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, a senior reporter who has covered AI, robotics, and more for eight years at The Verge.

Aug 6, 2020, 2:44 PM GMT+2 | □



Figure 5: [source](#)

Principles of Open Science

Open Science lifecycle

Roles in Open Science

Funders make open science part of the selection process, and conditions for grantees conducting research.

Publishers make open science part of the review process, and conditions for articles published in their journals.

Institutions make open science part of academic training, and part of the selection process for research positions and evaluation for advancement and promotion.

Societies make open science part of their awards, events, and scholarly norms.

Researchers enact open science in their work and advocate for broader adoption in their communities.

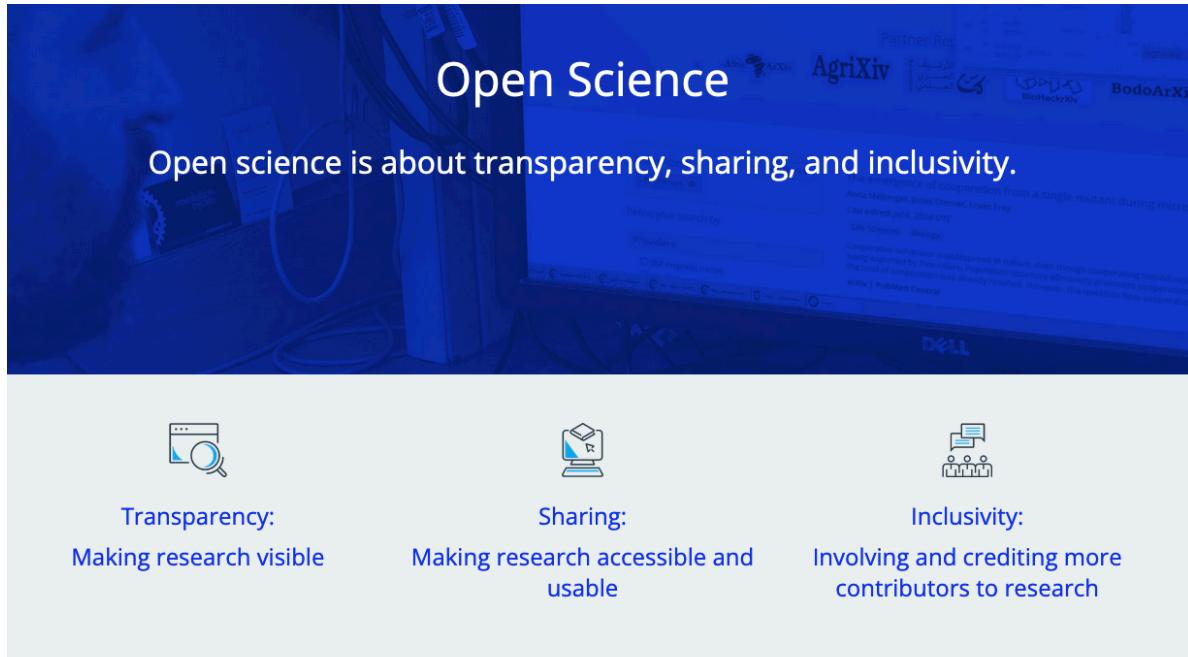
[Center for Open Science]

Skills to Pay the Bills

- Professors make professors
- How to get funding
- How to get published
- How to get cited
- How to give credit (citation)
- Research skills often *informally* transmitted

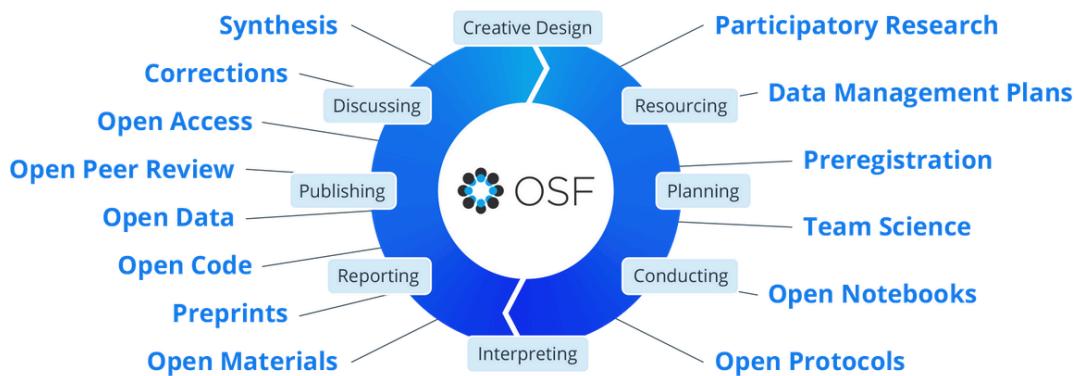


Figure 6: [source](#)



These principles aim to democratize access to research, promote equitable resource distribution, foster accountability and trustworthiness, accelerate self-correction, and improve rigor and reproducibility.

Figure 7: Center for Open Science



We advocate for lifecycle open science. There are open scholarship activities at every stage of the research lifecycle (see figure above) that individually and collectively contribute to improving science, with everyone playing a role:

Figure 8: Center for Open Science

Who profits from Open Science?

What is Open Science to you?

What do you find interesting, important, or attractive about Open Science?

<https://tinyurl.com/opnsci>

you pay my salary,
but you don't get access to my
work.



Figure 9: [source](#)

Learning outcomes

- * Open Science = Good science in a digitized world



- * Open Science impacts all steps in the research cycle
=> change in practices in planning, data collection, analysis, presentation, ...



- * Open Science = social change

→ makes it difficult (social hurdles)
→ it is possible, if we understand mechanisms
+ support each other



CC-BY 4.0 Heidi Seibold
@HeidiBoya

Implementing an open and reproducible workflow

1. version control
 2. project structure
 3. data
 4. methods
 5. code
 6. authoring
 7. publishing
-

Break

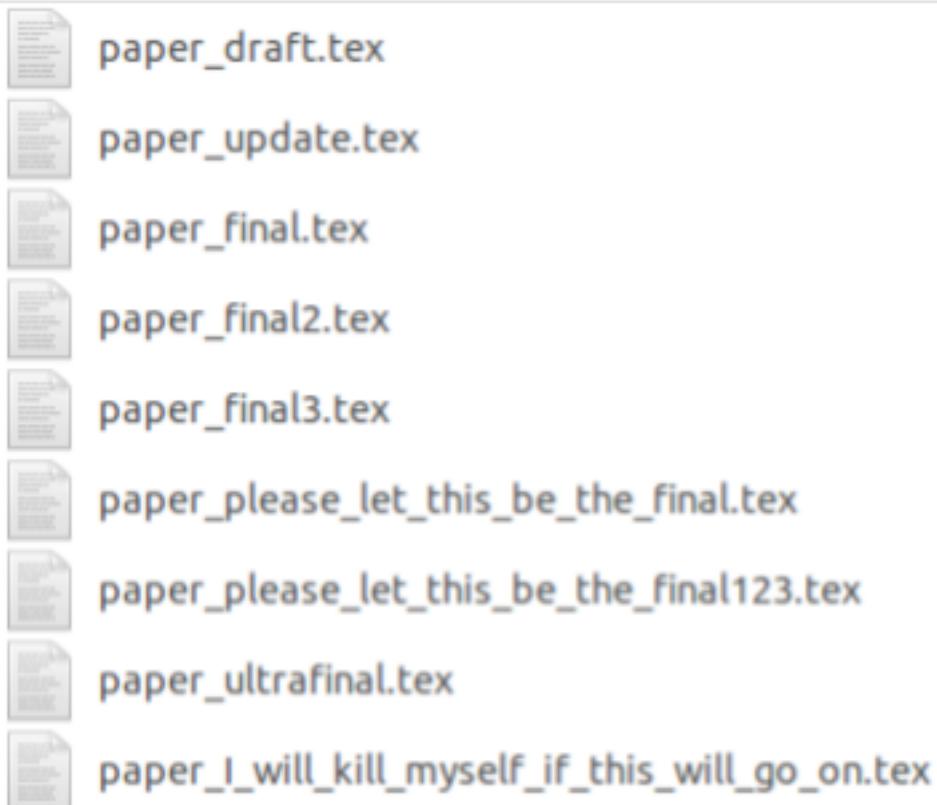
Taking notes

Networked notes

- logseq
 - Obsidian
-

Version control

Why use version control?



Dear colleagues,

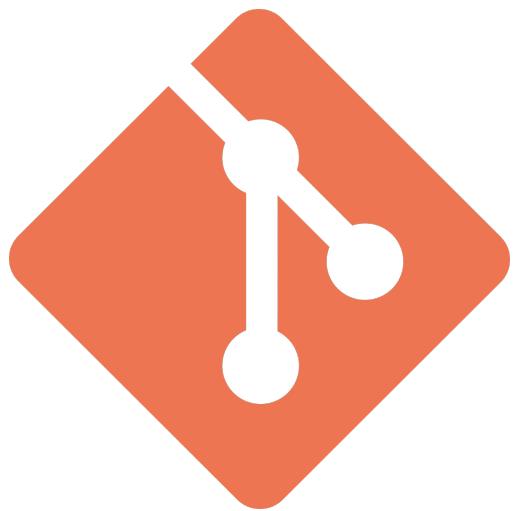
attached you find the first public version of the [REDACTED] protocol.
Please have a look and do comment. We can also meet to aggregate our reviews.

► 1 attachment: StudyProposal[REDACTED]Validation_V1_250918docx.docx

[source](#)

git and GitHub/GitLab

git software on your machine



```
git add scr/tests.py  
git commit -m 'add tests'  
git push
```

GitHub and GitLab services on a remote server





git commands

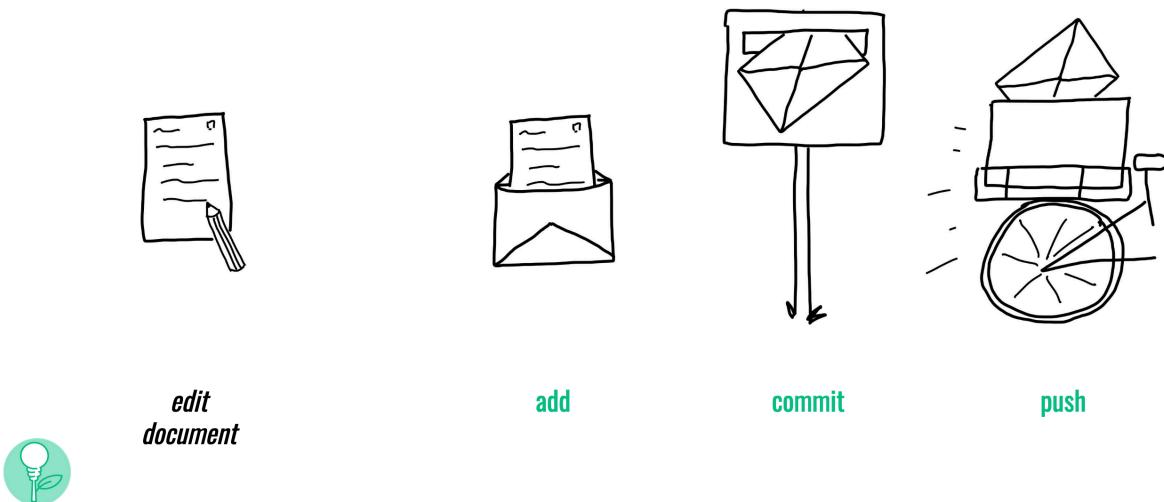
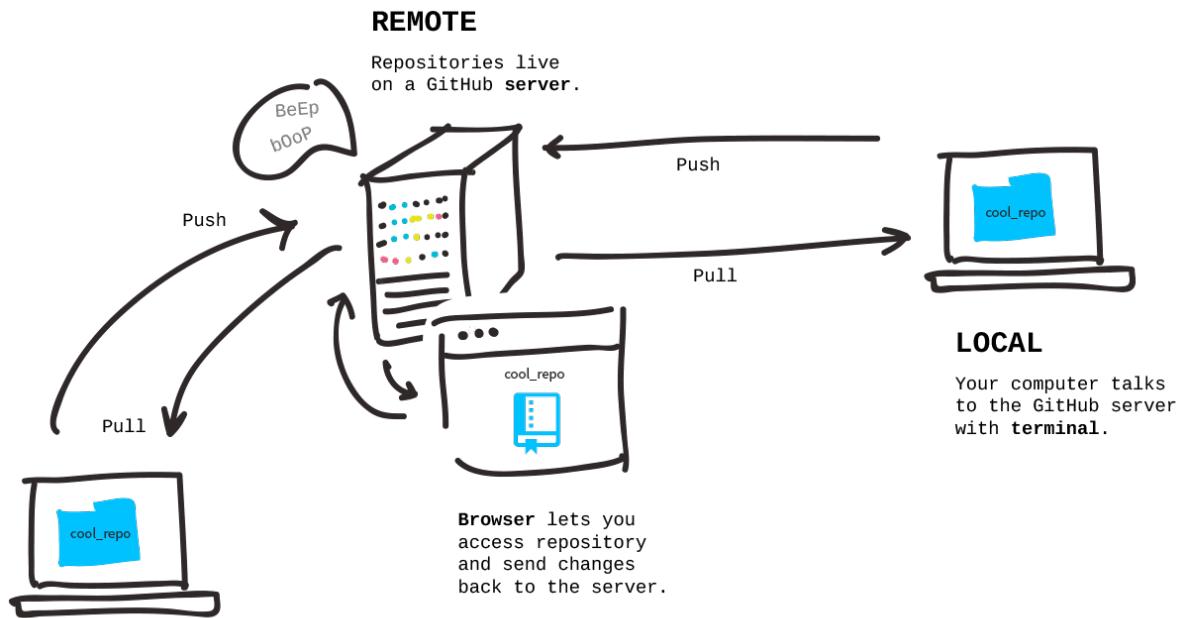


Figure 10: ([source](#))



LOCAL

Someone else's computer talks to the GitHub server.

Figure 11: ([source](#))

Collaborating using GitHub

GitHub workflow

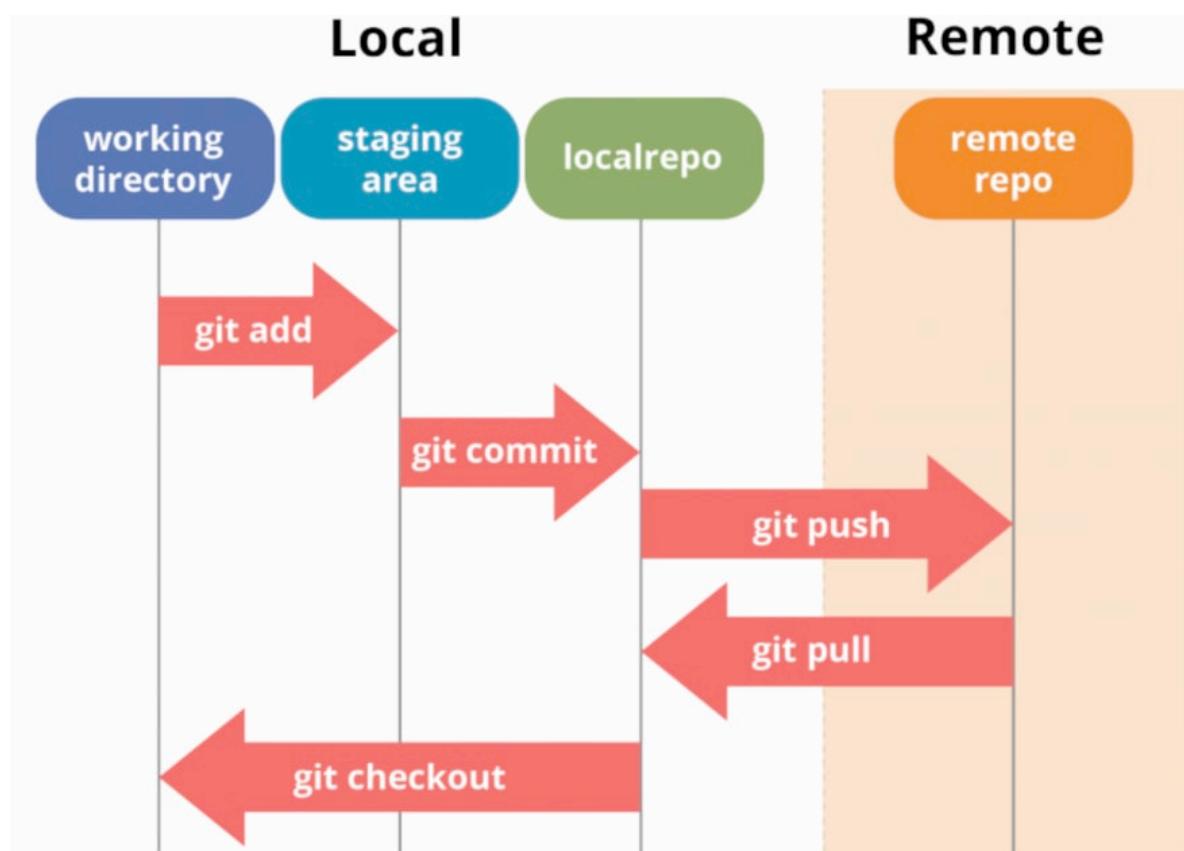


Figure 12: ([source](#))

How to set up a GitHub repository

create a repository on GitHub

1. (create [GitHub](#) account)
2. click on New (<https://github.com/new>)
3. specify repo name ¹
4. specify description
5. specify visibility: private or public
6. select Add a README file
7. specify licence ²

¹safe: lowercase alphabet characters

²good choice for many purposes: MIT license

Create a new repository

A repository contains all project files, including the revision history. Already have a project repository elsewhere? [Import a repository](#).

Required fields are marked with an asterisk ().*

Repository template

No template ▾

Start your repository with a template repository's contents.

Owner *

wuqui ▾

Repository name *

/ opensciencews

✓ opensciencews is available.

Great repository names are short and memorable. Need inspiration? How about [glowing-parakeet](#) ?

Description (optional)

Materials for the Open Science workshop at CAIS.

Public

Anyone on the internet can see this repository. You choose who can commit.

Private

You choose who can see and commit to this repository.

Initialize this repository with:

Add a README file

This is where you can write a long description for your project. [Learn more about READMEs](#).

Add .gitignore

.gitignore template: None ▾

Choose which files not to track from a list of templates. [Learn more about ignoring files](#).

Choose a license

License: MIT License ▾

clone the repository

go to the folder where you want your project to live



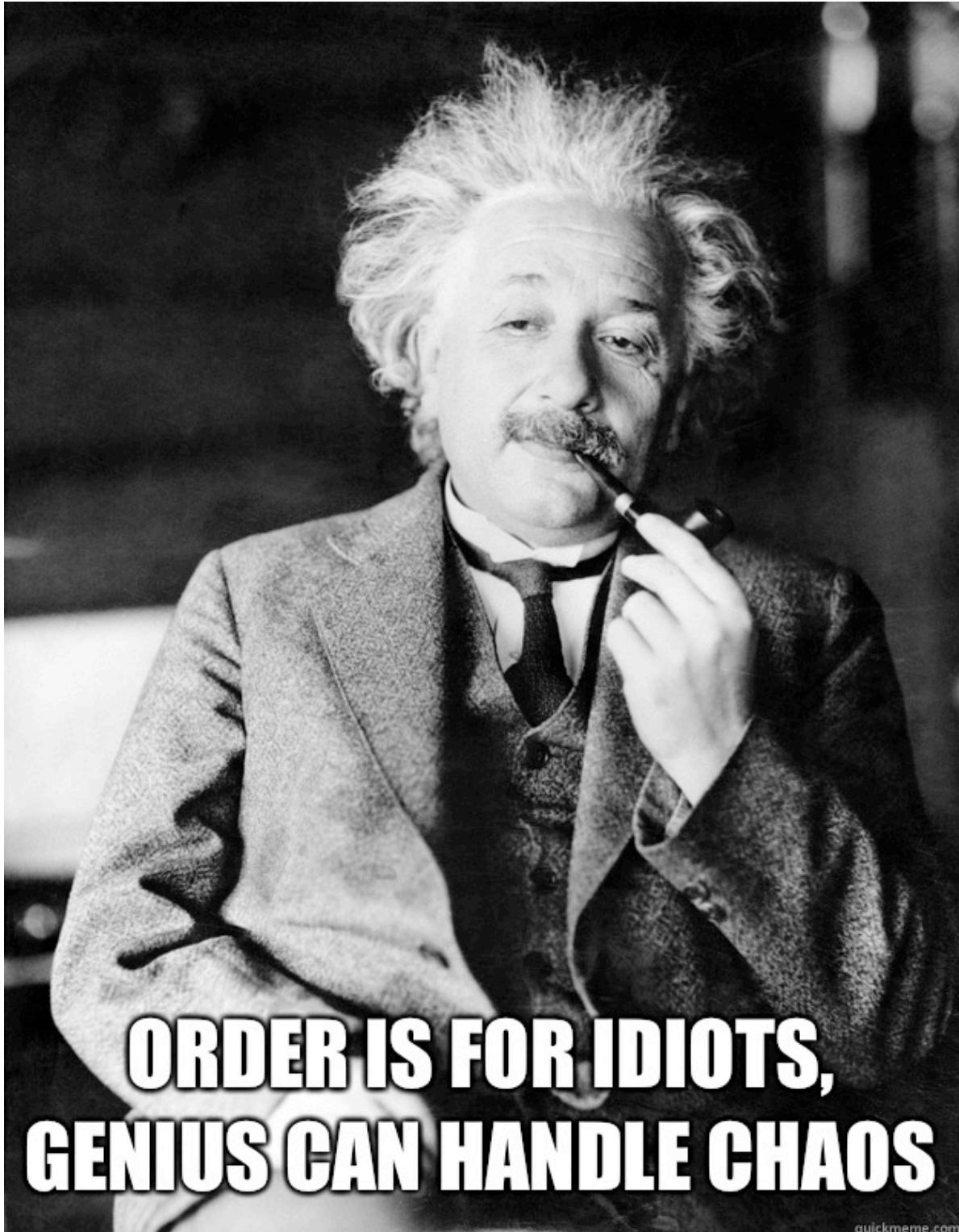
```
git clone https://github.com/wuqui/opensciws.git
```

add changes

commit changes

push changes

Project structure



Let's not pretend we're all geniuses ...

File names

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

File names should be:

- Machine readable
 - Human readable
 - Optional: Consistent
 - Optional: Play well with default ordering
-

File structure

```
.  
  analysis          <- all things data analysis  
    src             <- functions and other source files  
  comm  
    internal-comm   <- internal communication such as meeting notes  
    journal-comm    <- communication with the journal, e.g. peer review  
  data  
    data_clean      <- clean version of the data
```

```
data_raw           <- raw data (don't touch)
dissemination
  manuscripts
  posters
  presentations
documentation      <- documentation, e.g. data management plan
misc               <- miscellaneous files that don't fit elsewhere
```

tree DIR

cookie cutters

Practice: managing files and folders

Data

FAIR data

Types of data

- interviews
 - questionnaires
 - web
 - social media
-

What is FAIR DATA?

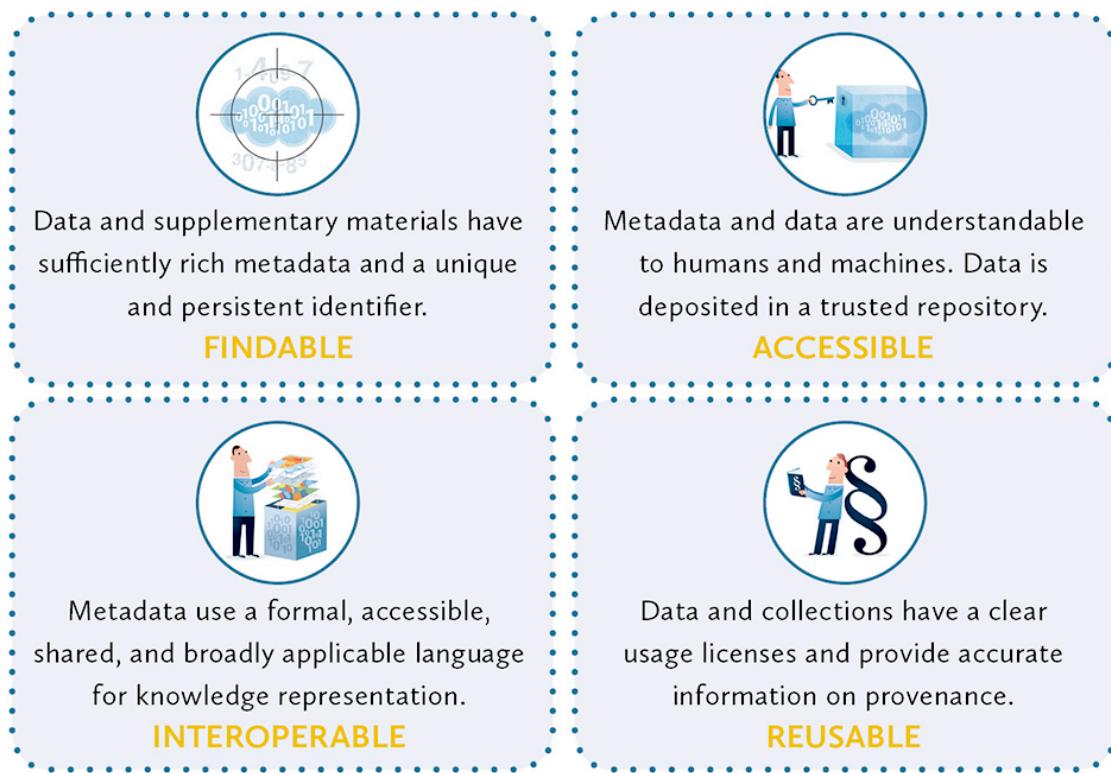


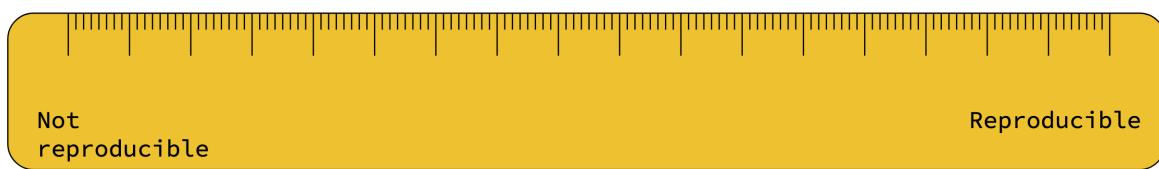
Figure 13: [source](#)

Code

Reproducibility et al.

		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

Figure 14: [The Turing Way](#)



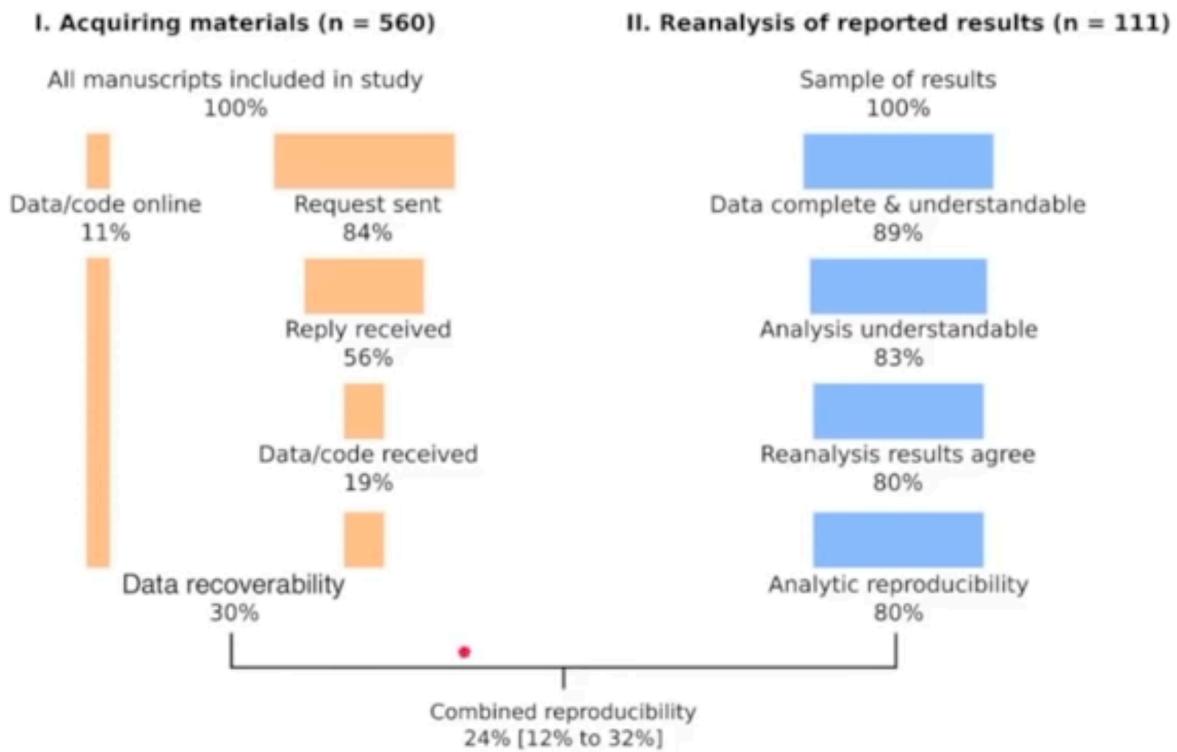


Figure 15: [source](#)

notebooks and literate programming

- Knuth
 - application
 - * R
 - * Python: nbdev
-

Licensing

Code

The screenshot shows a user interface for choosing a license. At the top, a question is asked: "Which of the following best describes your situation?". Three options are presented with icons:

- I need to work in a community.** (Icon: three people) Description: Use the [license preferred by the community](#) you're contributing to or depending on. Your project will fit right in. If you have a dependency that doesn't have a license, ask its maintainers to [add a license](#).
- I want it simple and permissive.** (Icon: up arrow) Description: The [MIT License](#) is short and to the point. It lets people do almost anything they want with your project, like making and distributing closed source versions. [Babel](#), [.NET](#), and [Rails](#) use the MIT License.
- I care about sharing improvements.** (Icon: double-headed arrow) Description: The [GNU GPLv3](#) also lets people do almost anything they want with your project, except distributing closed source versions. [Ansible](#), [Bash](#), and [GIMP](#) use the GNU GPLv3.

Below this, another section asks: "What if none of these work for me?". Three additional options are listed:

- My project isn't software.** Description: [There are licenses for that.](#)
- I want more choices.** Description: [More licenses are available.](#)
- I don't want to choose a license.** Description: [Here's what happens if you don't.](#)

Figure 16: <https://choosealicense.com>

Other materials

License Features

Your choices on this panel will update the other panels on this page.

Allow adaptations of your work to be shared?



Yes No Yes, as long as others share alike

Allow commercial uses of your work?



Yes No

Figure 17: <https://creativecommons.org/choose/>

Authoring

Authoring How can we organise our project from the beginning so that we can publish outputs in the end?

Publishing Where can I publish my work (platforms, research centers infrastructure, ...)?

Plain text

Quarto

- single source → multiple output formats
 - PDF for publication outlets

- blog
 - website
-

Publishing

How: How can we organise our project from the beginning so that we can publish outputs in the end? Where: Where can I publish my work (platforms, research centers infrastructure, ...)?

Pre-registration



Figure 18: [source](#)

Outlets

ArXiV preprints

Zenodo all kinds including data, code, preprints, etc.

GitHub and GitLab code, software

Open Science Framework all kinds including data, code, preprints, preregistration, etc.

Software Heritage archival of code (long-term)

Papers with Code code and data for and with papers, mostly Machine Learning

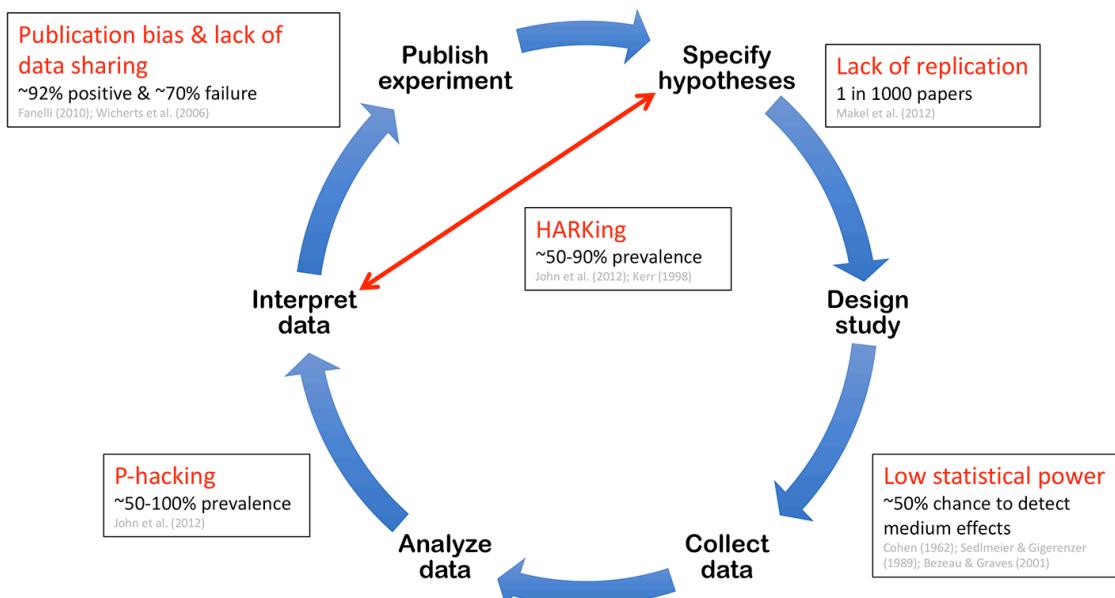


Figure 19: source

...

Resources

- DRA
- The Turing Way
- Data Carpentries