

# Open Science and Research Software Engineering

Workshop  
Center for Advanced Internet Studies (CAIS)

Quirin Würschinger

September 21, 2023

## Introduction

---

> `whoami`

Quirin Würschinger  
[q.wuerschinger@lmu.de](mailto: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
- 

## Workshop materials

**GitHub repository** <https://github.com/wuqui/opensciws>  
**slides** [https://wuqui.github.io/opensciws/opensciws\\_slides.html](https://wuqui.github.io/opensciws/opensciws_slides.html)  
**website version** [https://wuqui.github.io/opensciws/opensciws\\_website.html](https://wuqui.github.io/opensciws/opensciws_website.html)

---

## Open Open Science workshop

Focus on ...

- ask questions
  - discuss
  - apply and practice
  - collaborate
- 

## Time table

Topic	Start	End
intro	09:00	
Open Science principles		10:30
~		
version control	10:50	
project structure		12:00
~		

Topic	Start	End
code	13:15	
data and methods		14:40
~		
publishing	15:00	
authoring		15:25
recap, open issues, feedback	16:00	16:30

---

## **Addressing different backgrounds and goals**

---

### **Backgrounds and interests**

CAIS: Forschung zu Digitalisierung und Digitale Gesellschaft  
**research fields**

- education and pedagogy
- political science
- sociology
- communications studies
- ...

### **data and methods**

- qualitative interviews
  - text analysis
  - quantitative surveys
  - experimental designs
  - social media studies
  - ...
-

## **Survey: main interests**

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

## **Who are you?**

Please briefly introduce yourself ...

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

## **Open Science principles**

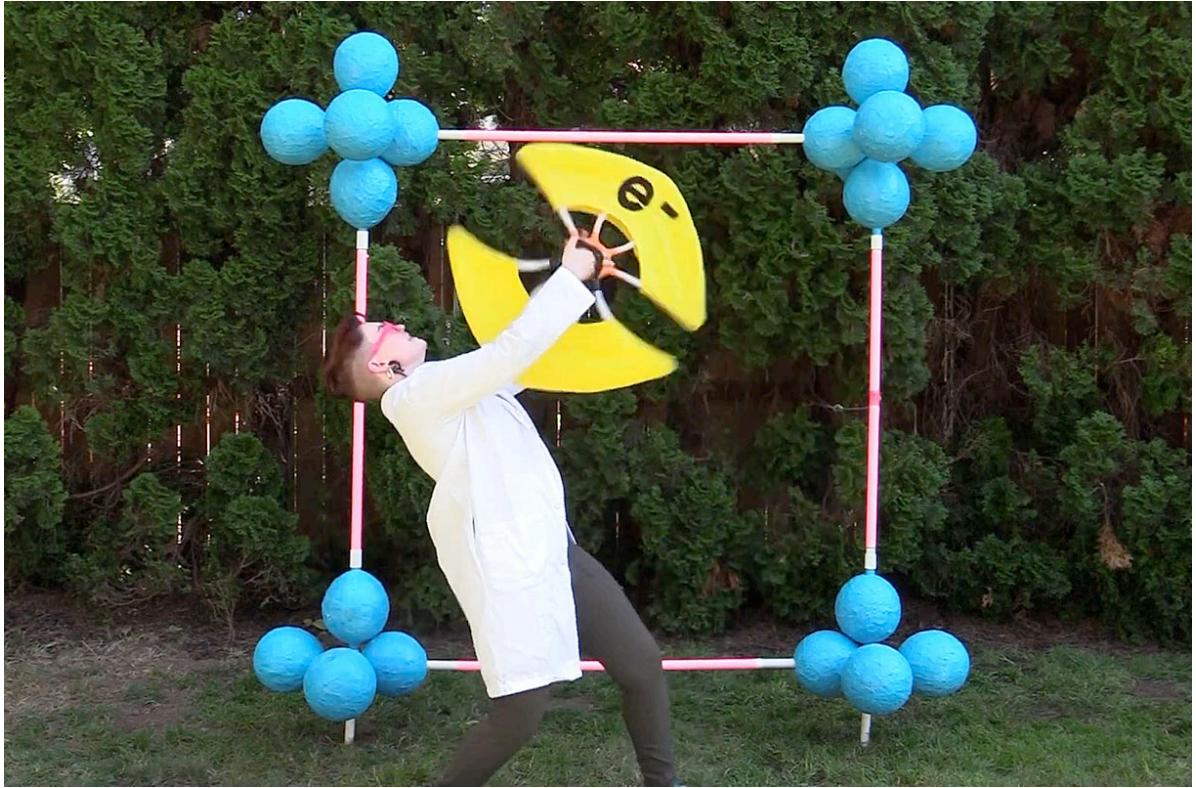
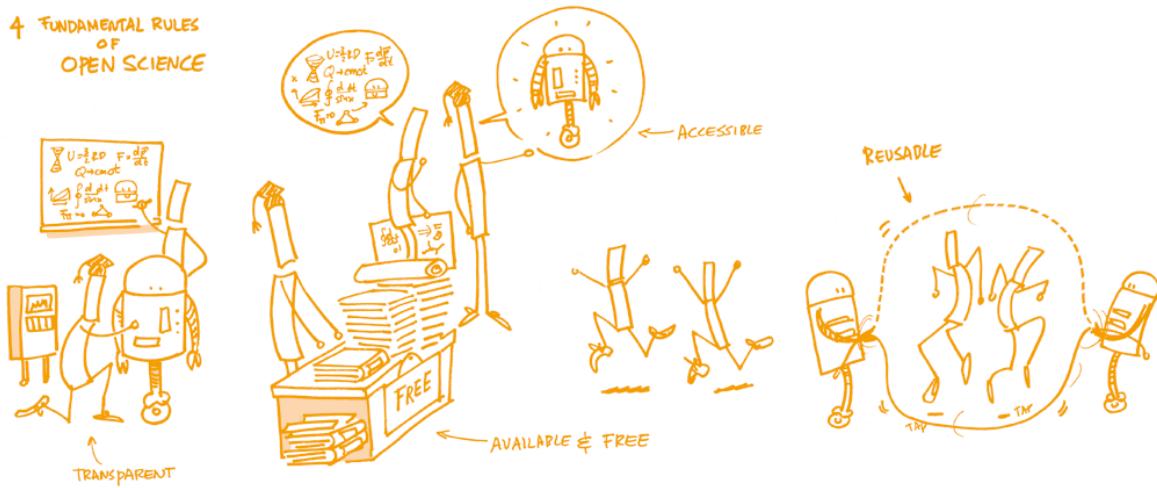


Figure 1: [Dance your PhD](#)

## What is Open Science?



---

## Why should we do Open Science?

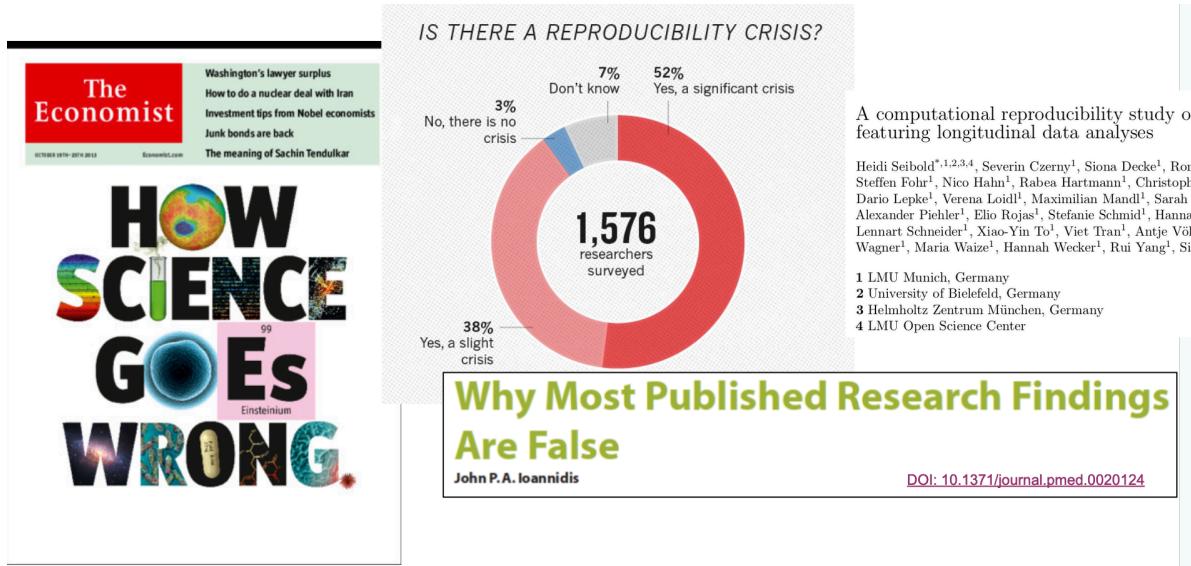
- dataset/sample size
  - effect sizes
  - selection/number of relationships
  - flexibility in design
  - financial interests
  - hype around topic/field
- 

What are the reasons why science can go wrong?

---

---

---



## Is this what we want?

Figure 2: [source](#)

## Principles of Open Science

09:20

## Open Science lifecycle

## WHAT FACTORS CONTRIBUTE TO IRREPRODUCIBLE RESEARCH?

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

● Always/often contribute   ● Sometimes contribute

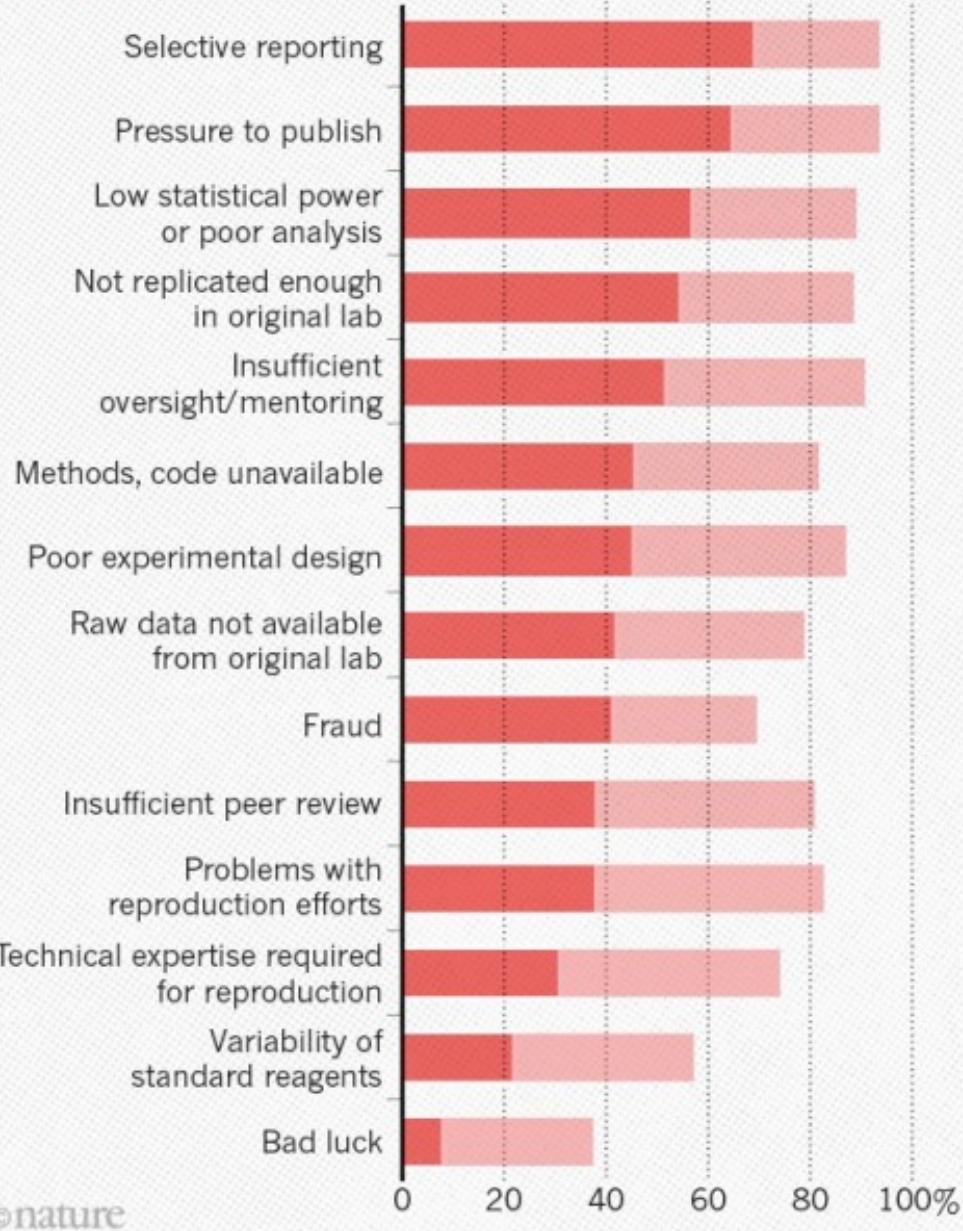


Figure 3: [source](#)

## 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

### 1 CHOOSE A POLITICAL PARTY

Republicans

Democrats

### 2 DEFINE TERMS

Which politicians do you want to include?

- Presidents
- Governors
- Senators
- Representatives

How do you want to measure economic performance?

- Employment
- Inflation
- GDP
- Stock prices

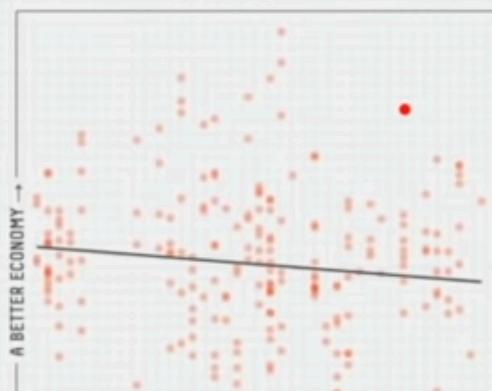
Other options

- Factor in power

<https://projects.fivethirtyeight.com/p-hacking/>

### 3 IS THERE A RELATIONSHIP?

Given how you've defined your terms, does the economy do better, worse or about the same when more Republicans are in office? Each dot below represents one month of data.



### 4 IS YOUR RESULT SIGNIFICANT?

If there were no connection between the economy and politics, what is the probability that you'd get results at least as strong as yours? That probability is your p-value, and by convention, you need a p-value of 0.05 or less to get published.



### Result: Almost

Your **0.07** p-value is close to the **0.05** threshold. Try tweaking your variables to see if you can push it over the line!

If you're interested in reading real (and more rigorous) studies on the connection between politics and the economy, see the work of Larry Bartels and Alan Blinder and Mark Watson.

Figure 4: [source](https://projects.fivethirtyeight.com/p-hacking/)

## *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

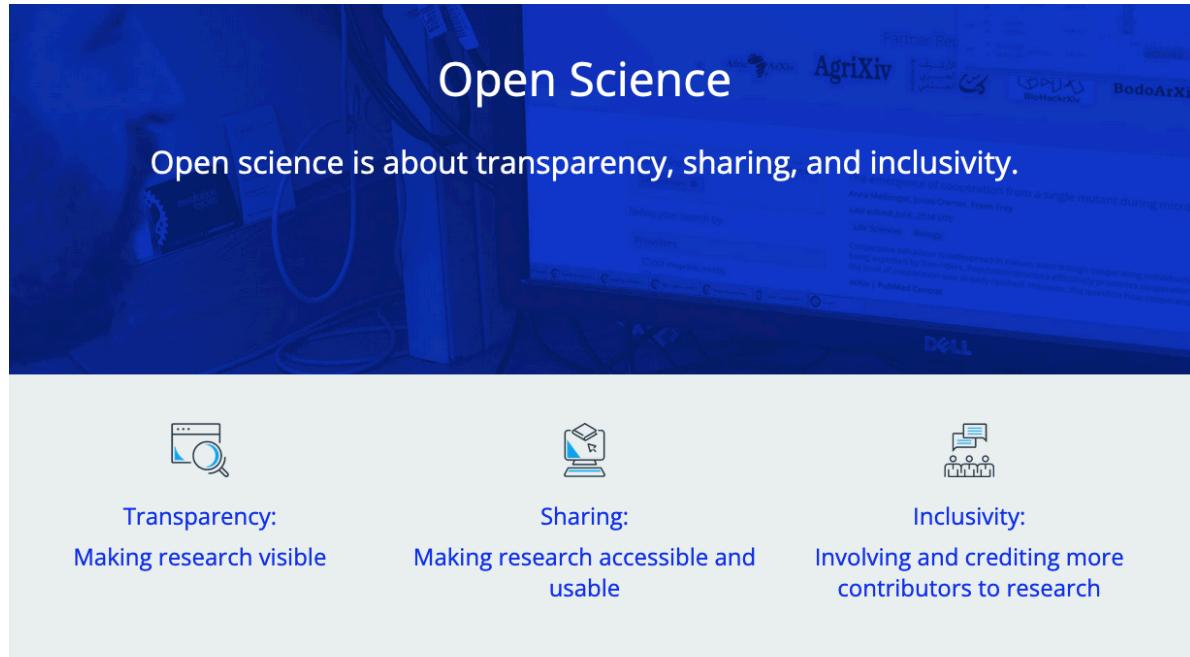
Figure 5: Richard McElreath: *Science as Amateur Software Development*

## *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

## 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)

---

## Who profits from Open Science?

---

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



Figure 9: [source](#)

## What is Open Science to you?

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

<https://tinyurl.com/opnsci>

09:50

---

## 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  
@HeidiBaya

---

## Implementing an open and reproducible workflow

1. version control
2. project structure
3. code

4. data and methods
  5. authoring
  6. publishing
- 
- 

## Version control

---

**Why use version control?**



`paper_draft.tex`



`paper_update.tex`



`paper_final.tex`



`paper_final2.tex`



`paper_final3.tex`



`paper_please_let_this_be_the_final.tex`



`paper_please_let_this_be_the_final123.tex`



`paper_ultrafinal.tex`



`paper_I_will_kill_myself_if_this_will_go_on.tex`

Dear colleagues,

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

▶ ⓘ 1 attachment: StudyProposal████████\_Validation\_V1\_250918docx.docx

[source](#)

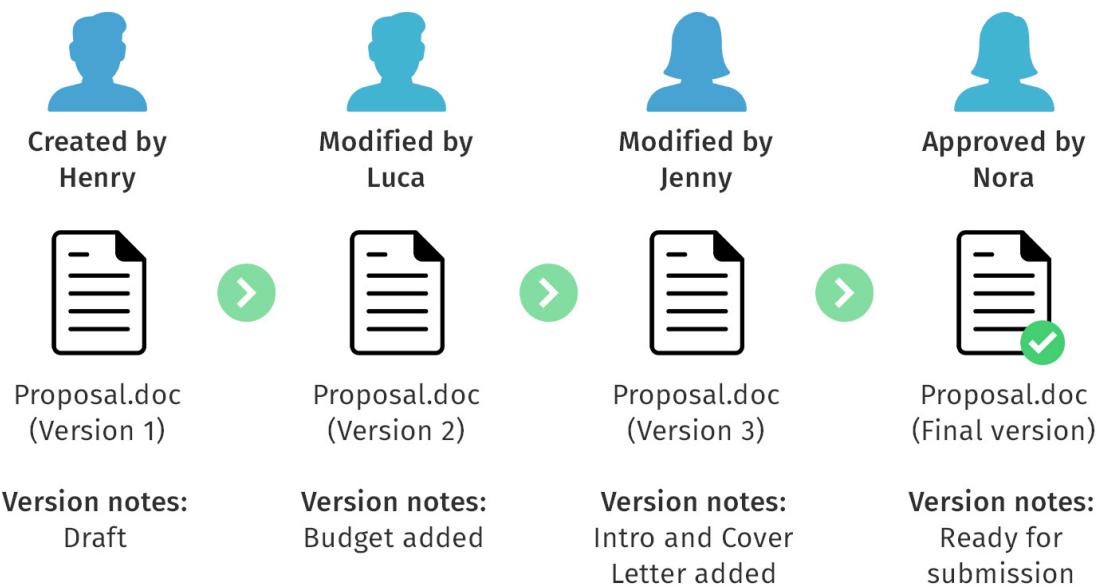


Figure 10: [source](#)

---

---

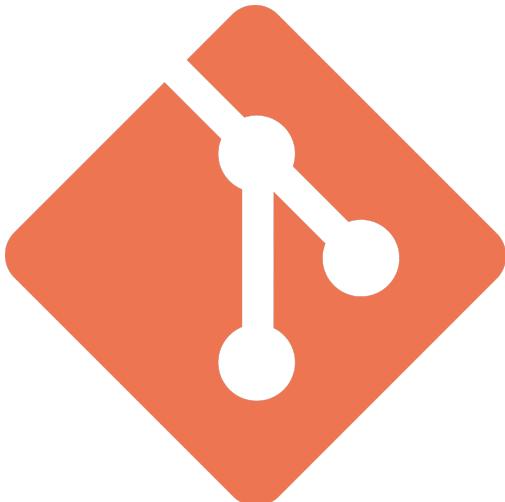
## git and GitHub/GitLab

**git** software on your machine

# BENEFITS OF DOCUMENT VERSION CONTROL



Figure 11: [source](#)



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

**GitHub and GitLab** services on a remote server



---

## Collaborating using GitHub

---

### git commands

---

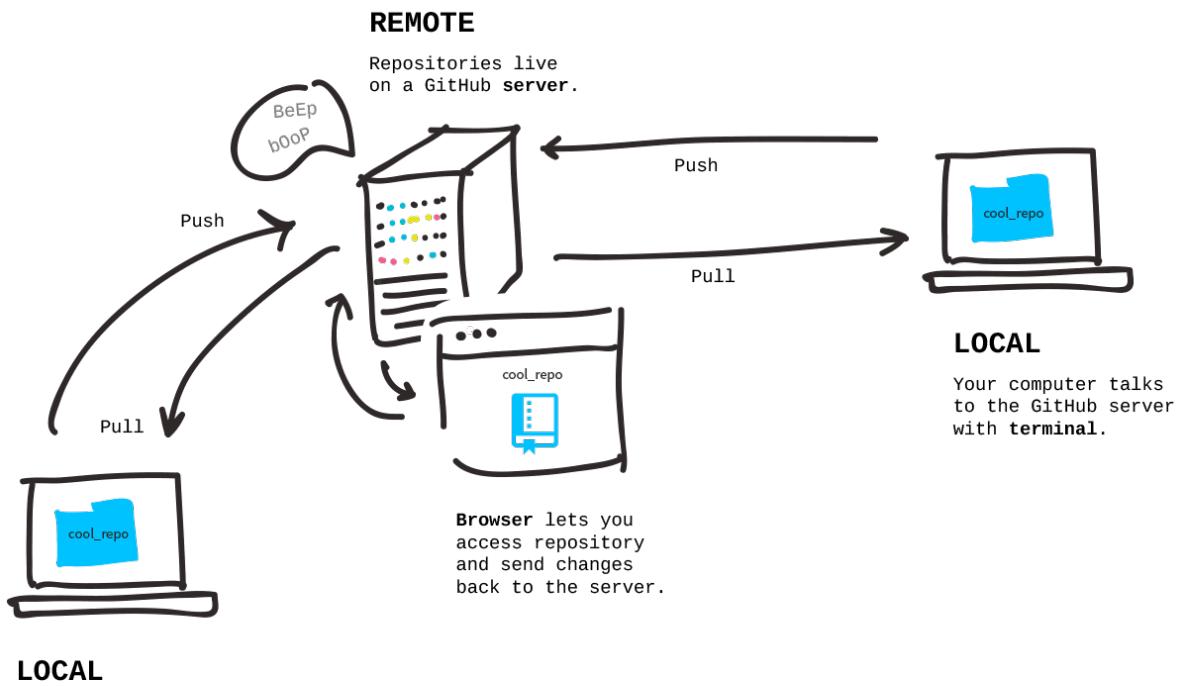


Figure 12: ([source](#))

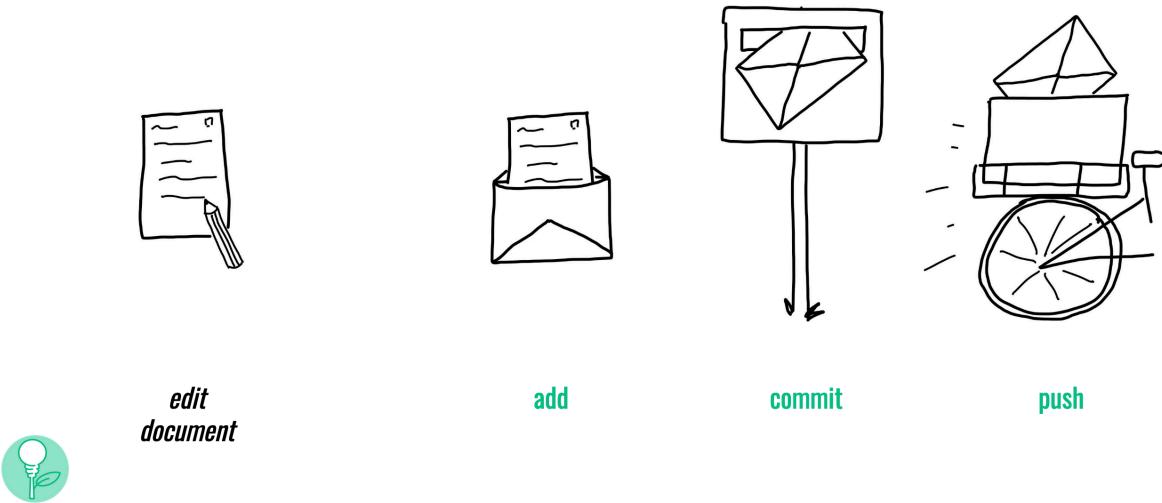


Figure 13: ([source](#))

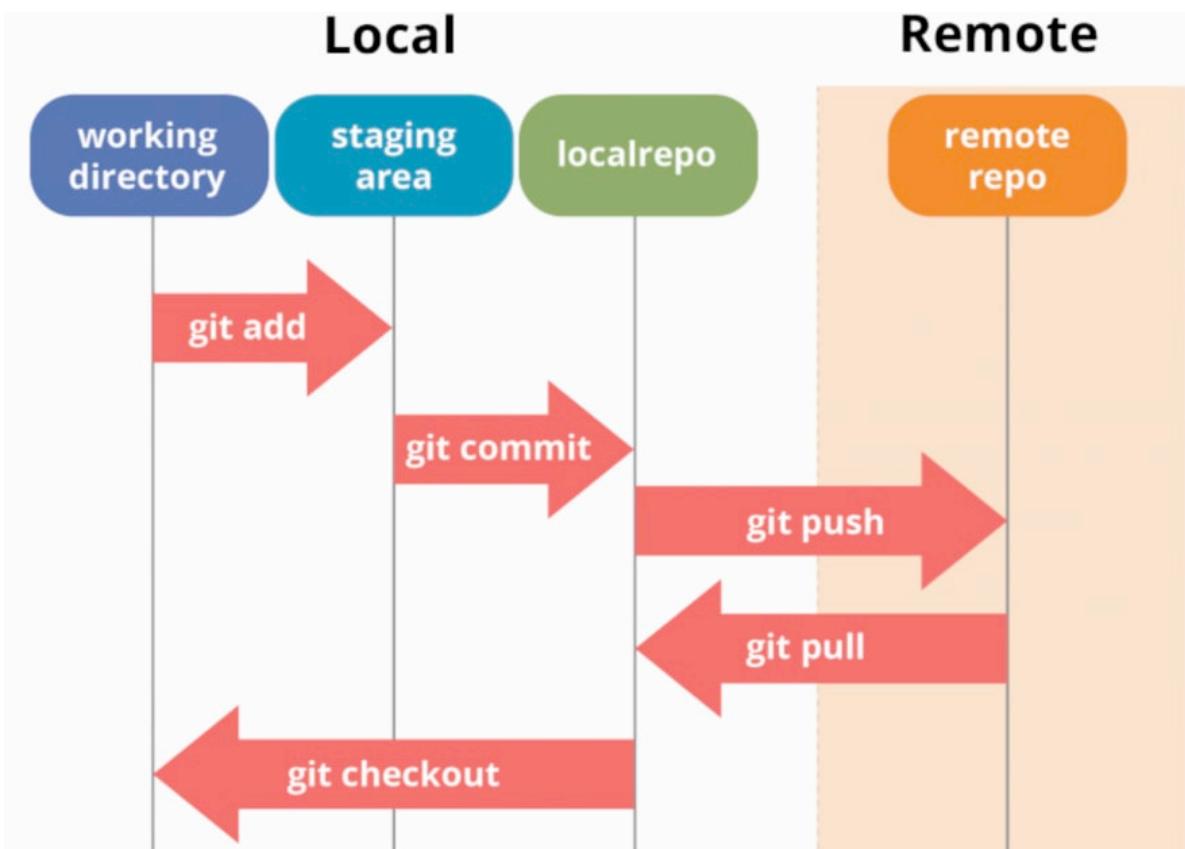


Figure 14: [\(source\)](#)

## GitHub workflow

---

### Example

The screenshot shows a GitHub pull request titled "Improve grammar in compendia.md (#1822)". The pull request has been merged into the "main" branch. The commit was made by "inwaves and malvikasharan" on April 24, 2021, and is verified. It has one parent commit, 328d54a, and a commit hash of 72039372a2803a4ec15764b2c01017039320f096. The diff shows two changes:

```
132 - In the future, the research compendium may even be the publication itself which is being peer reviewed  
      (rather than just peer reviewing the paper, why not review the entire research project).  
133 -  
133 + In the future, the research compendium may even be the publication itself allowing peer review of the entire  
      research project.
```

Figure 15: [source](#)

---

## How to set up a GitHub repository

---

**set up git**

**Installing git:** see [tutorial](#)

**Using git:**

- from the command line
- using a standalone GUI<sup>1</sup> tool; e.g.:
  - [GitKraken](#)
  - [GitHub Desktop](#)
- from within your editor/IDE<sup>2</sup>; e.g.:

---

<sup>1</sup>Graphical User Interface

<sup>2</sup>Integrated Development Environment

- RStudio
  - VSCode
- 

## **set up GitHub**

tutorial

- setting up git user information (name, password)
  - setting up GitHub authentication
  - setting and storing authentication ('token')
- 

## **create a repository on GitHub**

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

---

<sup>3</sup>safe: lowercase alphabet characters

<sup>4</sup>good 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 ▾

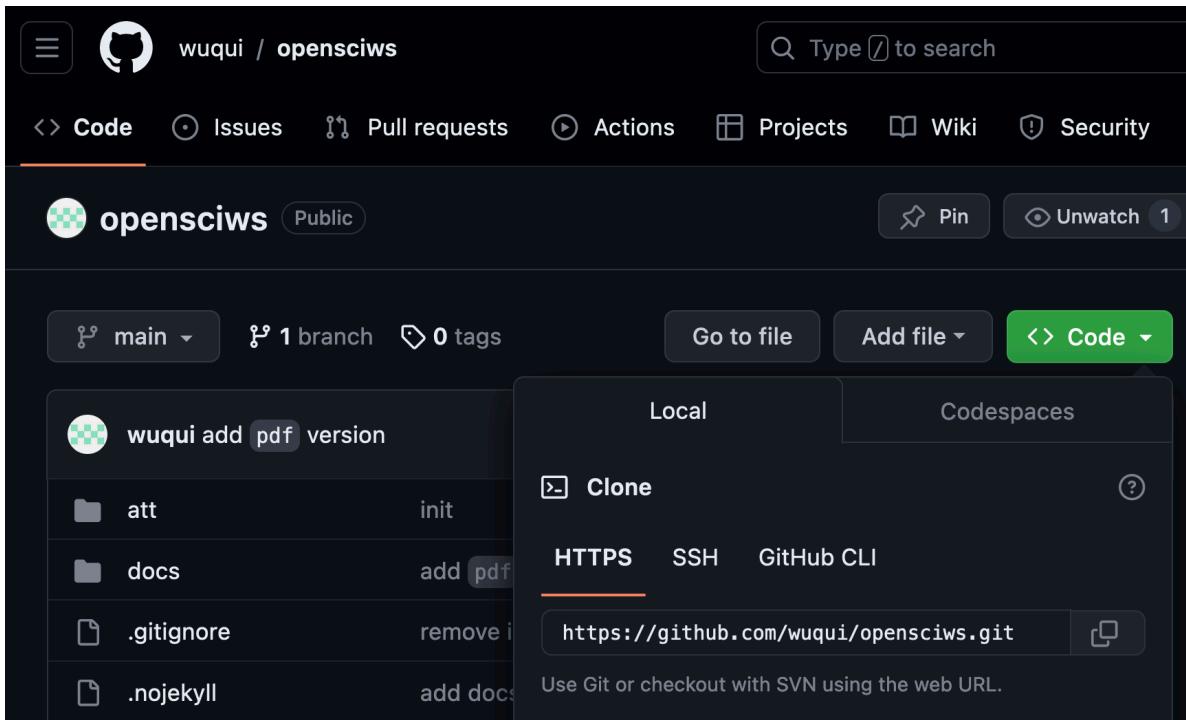
Choose which files not to track from a list of templates. [Learn more about ignoring files](#).

## Choose a license

License: MIT License ▾

## clone repositories

go to the folder where you want your project to live



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

---

## adding, committing, and pushing changes

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

---

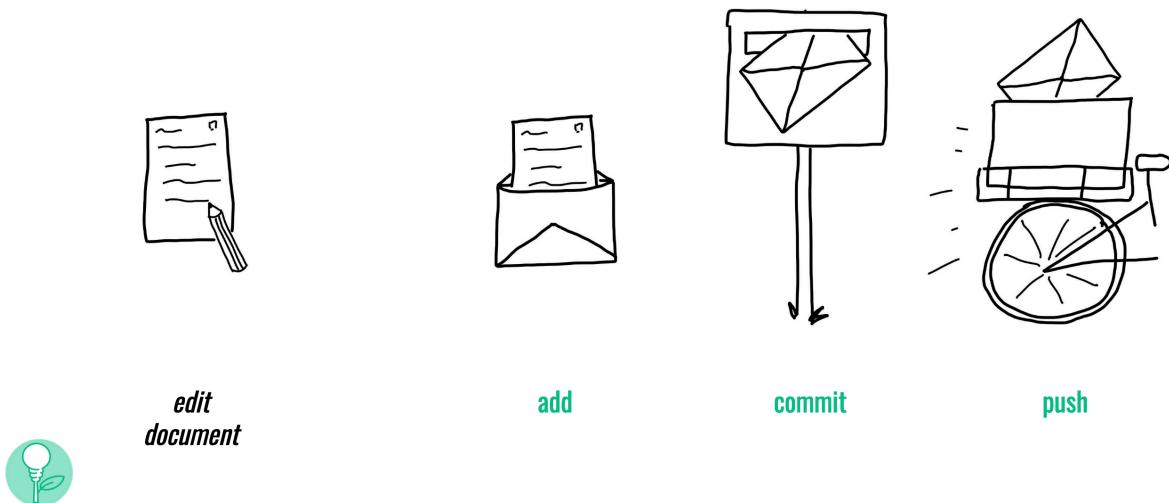
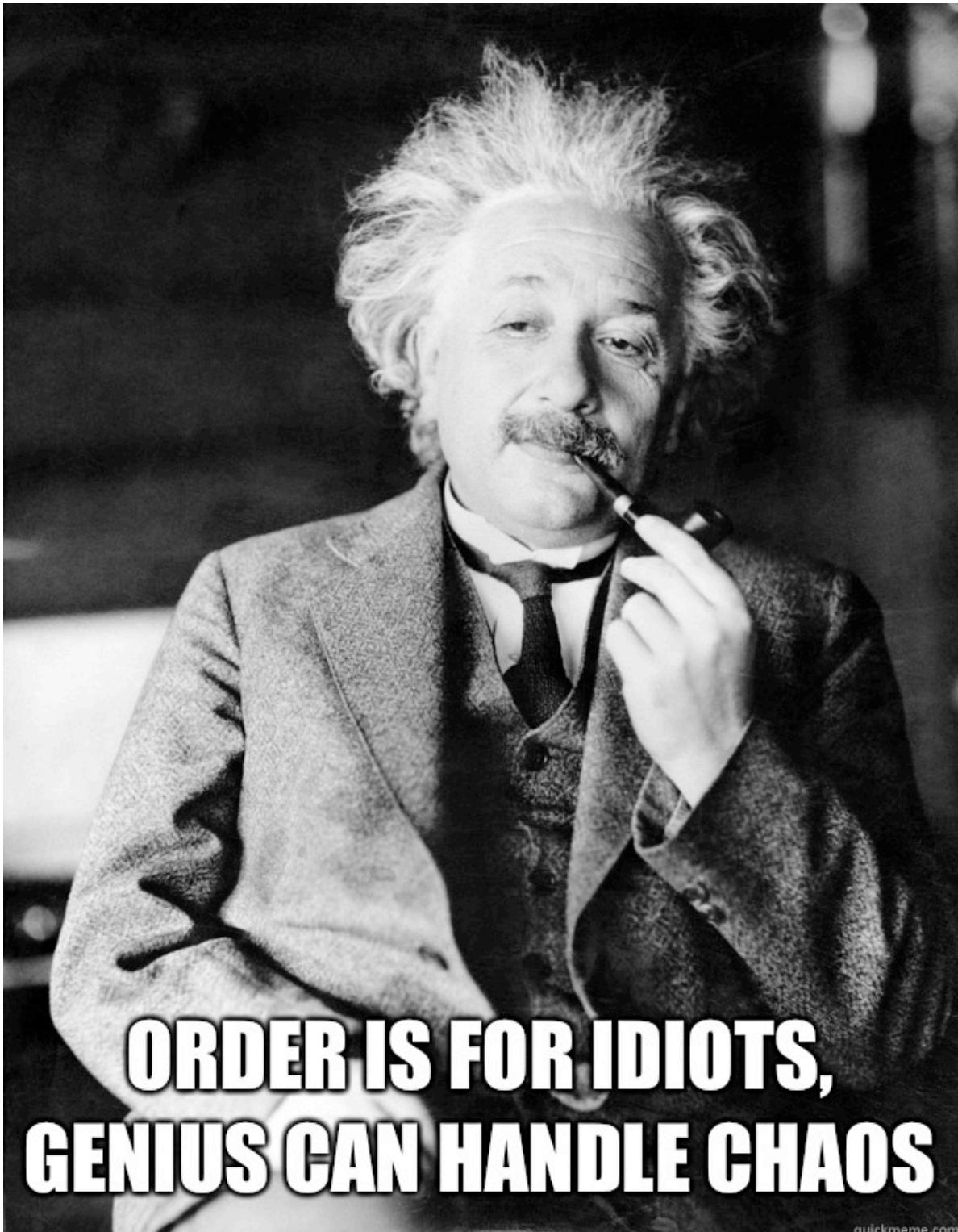


Figure 16: (source)

## 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
  - consistent
  - optional: play well with default ordering (e.g. include timestamps)
- 

## 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
```

---

## Practice: project management

You have until 11:50 h to work on either ...

1. developing a project structure for your needs from scratch
2. refactoring/cleaning an existing project<sup>5</sup>

Optionally: set up version control via git/GitHub for this project.

---

---

## Code

---

## Reproducibility

---

## Reproducibility (crisis)

---

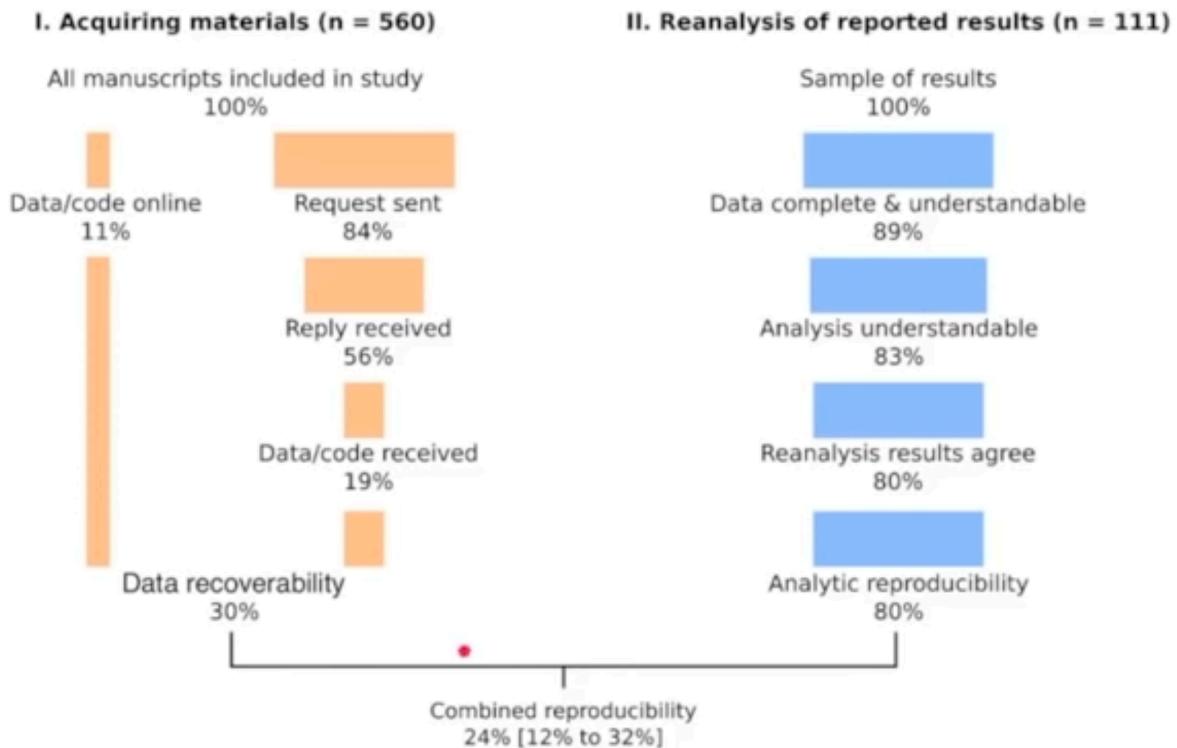


Figure 17: [source](#)

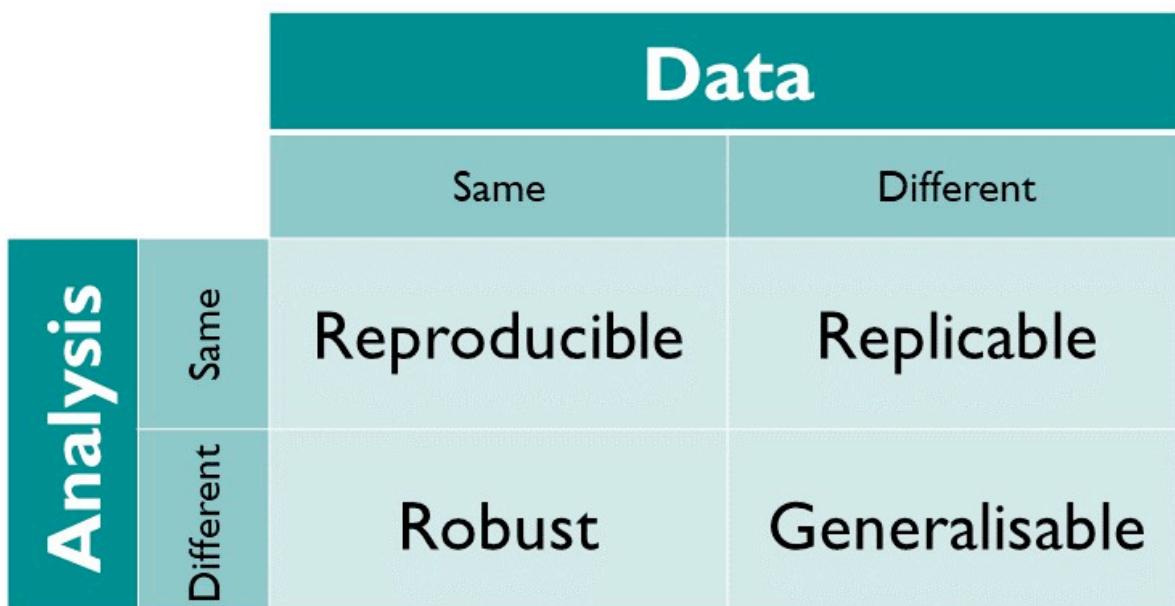
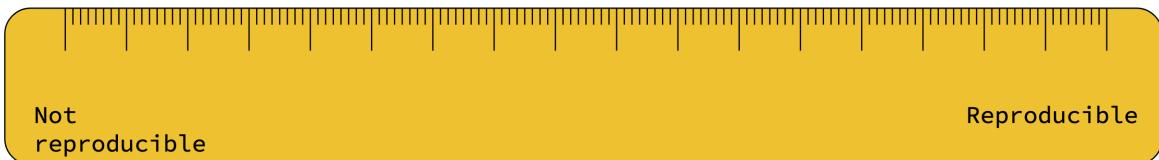


Figure 18: [The Turing Way](#)

## Reproducibility et al.



interesting for student projects

---

## The quality of tools

SCIENCE / TECH / MICROSOFT

# 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 | □

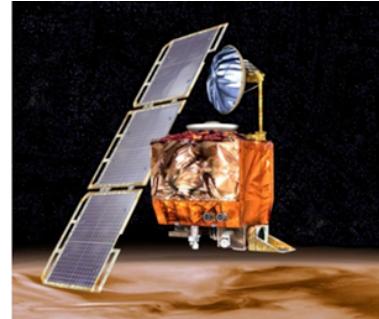


Figure 19: [source](#)

---

LISA GROSSMAN 11.10.10 7:00 AM

# NOV. 10, 1999: METRIC MATH MISTAKE MUFFED MARS METEOROLOGY MISSION



The **\$125 million satellite** was supposed to be the first weather observer on another world. But as it approached the red planet to slip into a stable orbit Sept. 23, the orbiter vanished. Scientists realized quickly it was gone for good. “It was pretty clear that morning, within half-an-hour, that the spacecraft had more or less **hit the top of the atmosphere and burned up**,” recalled NASA engineer Richard Cook, who was project manager for Mars exploration projects at the time.

A NASA review board found that the problem was in the software controlling the orbiter’s thrusters. **The software calculated the force the thrusters needed to exert in pounds of force. A separate piece of software took in the data assuming it was in the metric unit: newtons.**

<https://www.wired.com/2010/11/1110mars-climate-observer-report/>

Figure 20: [source](#)

## Testing code

### Why we should test code

---

not all our projects have that high stakes

### Professional testing

#### math

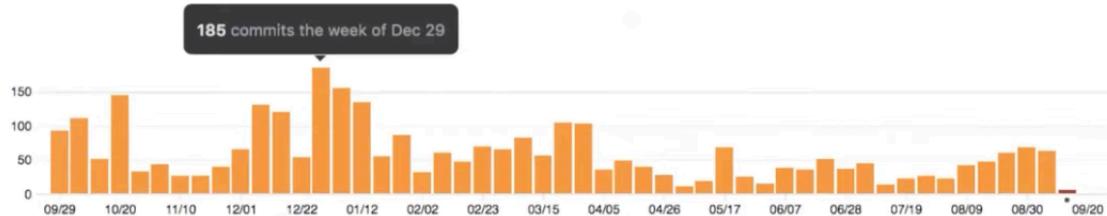
The Stan Math Library is a C++ template library for automatic differentiation of any order using forward, reverse, and mixed modes. It includes a range of built-in functions for probabilistic modeling, linear algebra, and equation solving.



math    automatic-differentiation    stan    stan-math-library  
C++    BSD-3-Clause    127    449    214 (20 issues need help)    23  
Updated 5 minutes ago

3.6 MB of library code

7.6 MB of test code



<https://github.com/stan-dev/math>

Figure 21: [source](#)

same is true of industry

---

<sup>5</sup>make a backup first

## Types of tests

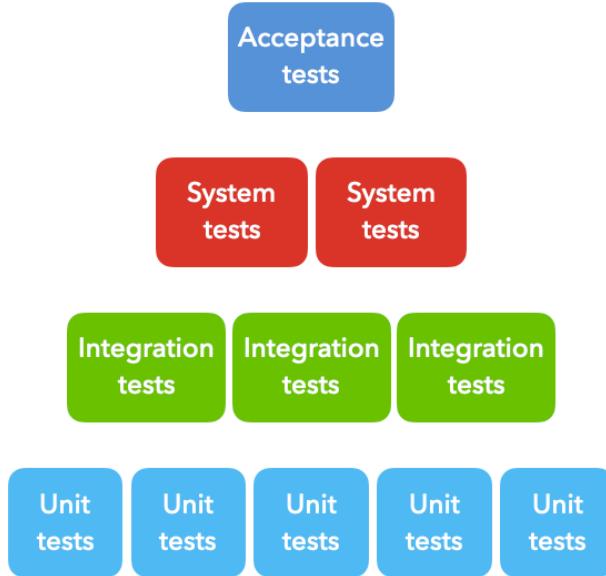


Figure 22: [source](#)

## Testing frameworks

- Python: `pytest`
- R: `testthat`

## Analogy

- during the process of manufacturing a ballpoint pen, the cap, the body, the tail, the ink cartridge and the ballpoint are produced separately and **unit tested** separately.
- When two or more units are ready, they are assembled and **integration testing** is performed, for example a test to check the cap fits on the body.
- When the complete pen is integrated, **system testing** is performed to check it can be used to write like any pen should.
- **Acceptance testing** could be a check to ensure the pen is the colour the customer ordered.

[source](#)

## Testing example

using pytest for Python

### Python

```
# test_assert_examples.py

def test_uppercase():
    assert "loud noises".upper() == "LOUD NOISES"

def test_reversed():
    assert list(reversed([1, 2, 3, 4])) == [4, 3, 2, 1]

def test_some_primes():
    assert 37 in {
        num
        for num in range(2, 50)
        if not any(num % div == 0 for div in range(2, num))
    }
```

---

## Documenting code

### Literate programming

- 'Literate programming is a methodology that combines a programming language with a **documentation language**,
- thereby making programs **more robust, more portable, more easily maintained**,
- and arguably **more fun** to write than programs that are written only in a high-level language.
- The main idea is to treat a program as a piece of literature, **addressed to human beings** rather than to a computer.
- The program is also viewed as a **hypertext document**, rather like the World Wide Web. (Indeed, I used the word WEB for this purpose long before CERN grabbed it!)

[Donald Knuth](#)

psychological benefit: conversation - helps reasoning - more fun (human) - → ChatBots

---

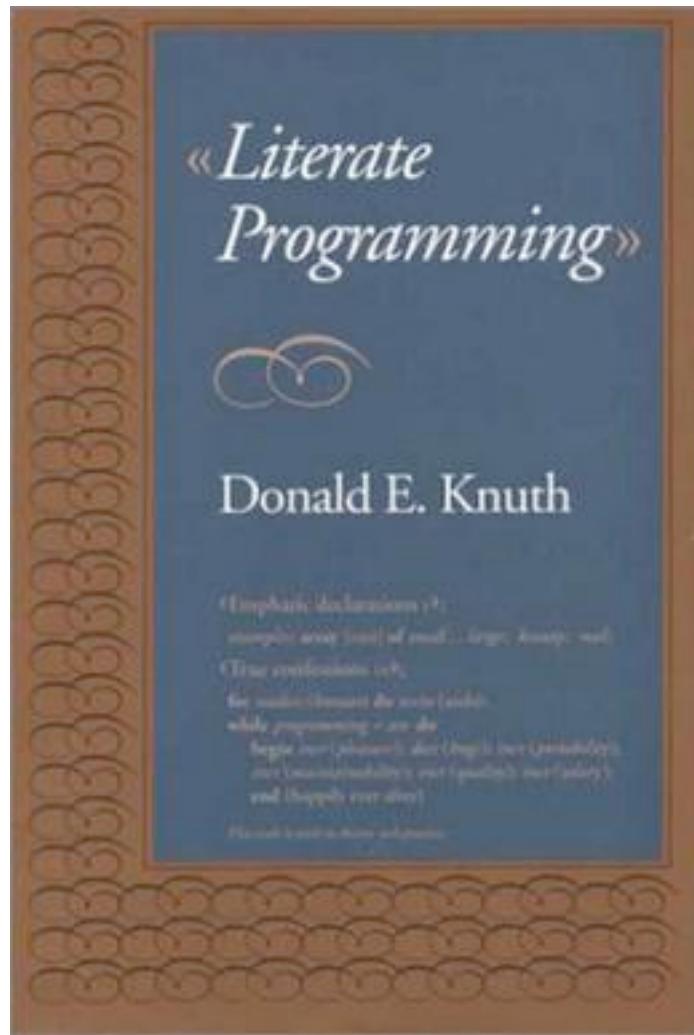
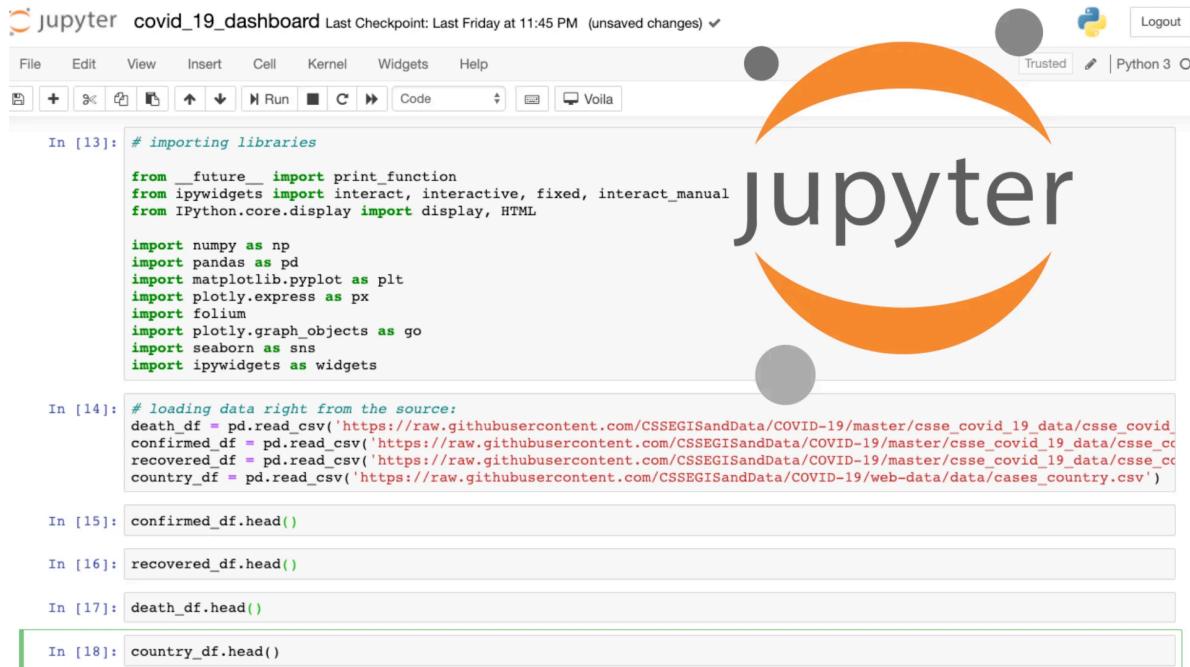


Figure 23: [source](#)

## Notebooks

### Python



The screenshot shows a Jupyter Notebook interface with the title "jupyter covid\_19\_dashboard". The notebook has a "Trusted" status and is running in "Python 3". The code cells contain Python code for importing libraries and loading COVID-19 data from CSV files. The interface includes a toolbar with file operations like File, Edit, View, Insert, Cell, Kernel, Widgets, Help, and a "Run" button. A large orange "jupyter" logo watermark is overlaid on the right side of the notebook area.

```
In [13]: # importing libraries
from __future__ import print_function
from ipywidgets import interact, interactive, fixed, interact_manual
from IPython.core.display import display, HTML

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import plotly.express as px
import folium
import plotly.graph_objects as go
import seaborn as sns
import ipywidgets as widgets

In [14]: # loading data right from the source:
death_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_deaths.csv')
confirmed_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_confirmed.csv')
recovered_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_recovered.csv')
country_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/web-data/data/cases_country.csv')

In [15]: confirmed_df.head()

In [16]: recovered_df.head()

In [17]: death_df.head()

In [18]: country_df.head()
```

Figure 24: [source](#)

### R

→ both work with [Quarto](#)

- who uses notebooks?
- which ones?
- good for novices & experts

---

### Example using nbdev for Python

Programming a deck of cards: [https://github.com/fastai/nbdev\\_cards/](https://github.com/fastai/nbdev_cards/)

```
---
```

```
title: "Components of a Quarto document"
output: html_document
date: "2022-08-24"
execute:
  echo: false
---
```

## Human-readable text

This is a Quarto document based on an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. When you click the **Render** button a document will be generated that includes both content as well as the output of any embedded code chunks within the document. You can embed an code chunk like this:

```
{r cars}
summary(cars)
```

## Including code chunks

You can embed plots, for example:

```
{r pressure}
#| echo = FALSE
plot(pressure)
```

Figure 25: [source](#)

```
In [ ]: #| export
suits = ["♣", "♦", "♥", "♠"]
ranks = [None, "A"] + [str(x) for x in range(2,11)] + ["J", "Q", "K"]
```

We will be using numbers to represent playing card clubs and ranks. These are the suits:

```
In [ ]: suits
```

```
Out[ ]: ['♣', '♦', '♥', '♠']
```

For instance the suit at index `0`:

```
In [ ]: suits[0]
```

```
Out[ ]: '♣'
```

These are the ranks:

```
In [ ]: ranks
```

```
Out[ ]: [None, 'A', '2', '3', '4', '5', '6', '7', '8', '9', '10', 'J', 'Q', 'K']
```

For instance the rank at index `1` (note that there isn't a playing card at position `0`, since we want the ranks to match the indices where possible):

---

## Literate testing with nbdev

For instance, here's a test of equality...

```
In [ ]: test_eq(Card(suit=1, rank=3), Card(suit=1, rank=3))
```

```
In [ ]: #| hide  
test_ne(Card(suit=2, rank=3), Card(suit=1, rank=3))  
test_ne(Card(suit=1, rank=2), Card(suit=1, rank=3))
```

...and a test of < ...

```
In [ ]: assert Card(suit=1, rank=3)
```

...and finally of > :

```
In [ ]: assert Card(suit=3, rank=3)>Card(suit=2, rank=3)  
assert not Card(suit=1, rank=3)>Card(suit=2, rank=3)
```

compare with pytest: much easier/natural

---

## Additional benefits of nbdev

- simple, integrated testing
- continuous integration
- dependency management
- publishing code for PyPI and conda
- publishing documentation via Quarto
- good for novices & experts
- covers about 80% of programming setup for free
- more about Quarto later

---

## R: Quarto and RMarkdown

```
---
```

```
title: "Components of a Quarto document"
output: html_document
date: "2022-08-24"
execute:
  echo: false
---
```

### Human-readable text

This is a Quarto document based on an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. When you click the **Render** button a document will be generated that includes both content as well as the output of any embedded code chunks within the document. You can embed an code chunk like this:

```
{r cars}
summary(cars)
```

### Including code chunks

You can embed plots, for example:

```
{r pressure}
#| echo = FALSE
plot(pressure)
```

benefit: visual editor

---

## Licensing

### Code

often appropriate: [MIT license](#)

## { Which of the following best describes your situation? }



### I need to work in a community.

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.

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.

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.

## { What if none of these work for me? }

### My project isn't software.

[There are licenses for that.](#)

### I want more choices.

[More licenses are available.](#)

### I don't want to choose a license.

[Here's what happens if you don't.](#)

Figure 26: <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 27: <https://creativecommons.org/choose/>

---

---

## Data and methods

---

### Diversity in data and methods

CAIS: Forschung zu Digitalisierung und Digitale Gesellschaft

#### research fields

- education and pedagogy
- political science
- sociology
- communications studies
- ...

## data and methods

- qualitative interviews
  - text analysis
  - quantitative surveys
  - experimental designs
  - social media studies
  - ...
- 

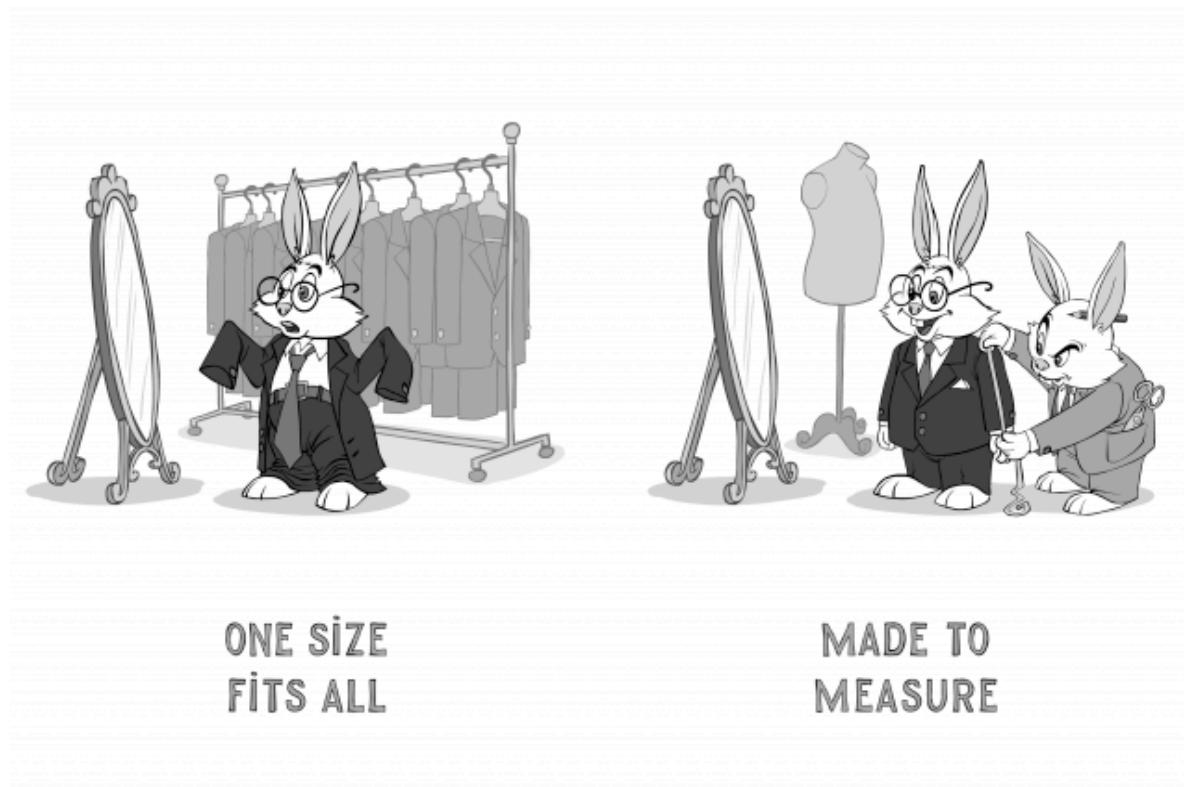


Figure 28: [source](#)

---

## Reasons to share your data

- To allow the possibility to fully **reproduce** a scientific study.

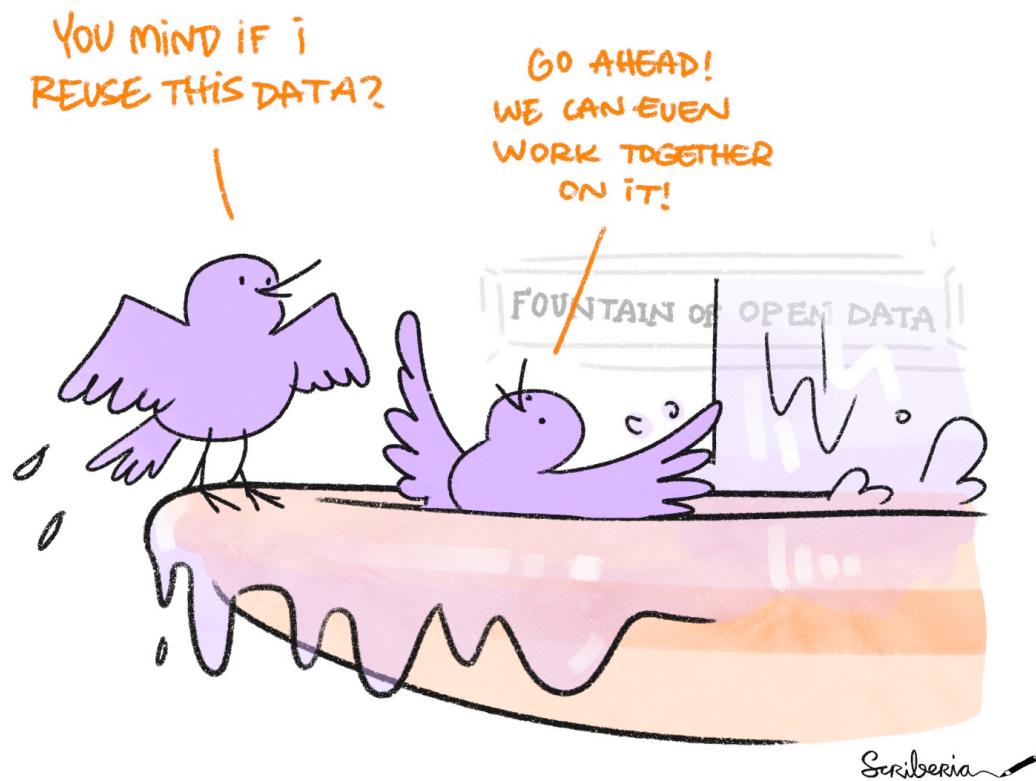


Figure 29: [source](#)

- To prevent duplicate efforts and **speed up scientific progress**. Large amounts of research funds and careers of researchers can be wasted by only sharing a small part of research in the form of publications.
  - To facilitate **collaboration** and increase the impact and quality of scientific research.
  - To make results of research openly available as a **public good**, since research is often publicly funded.
- 

## FAIR data

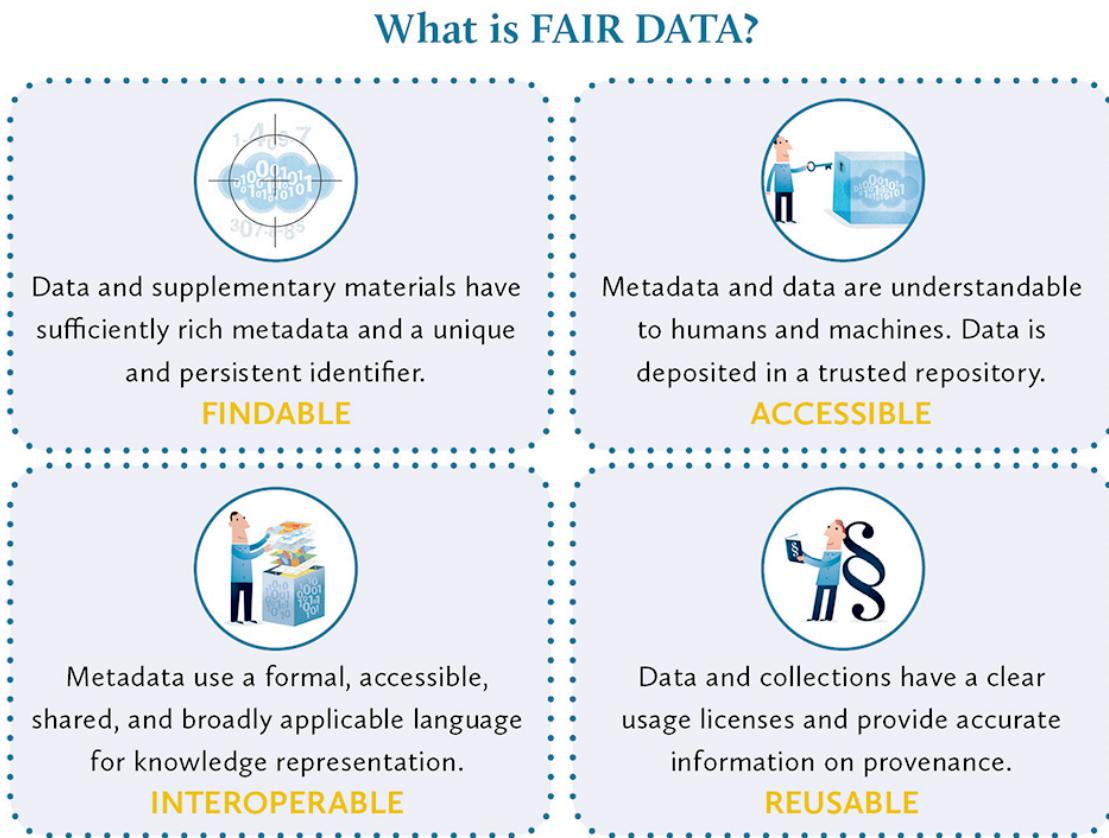


Figure 30: [source](#)

## How to share your data

### Turing Way tutorial

- Step 1: Select **what data** you want to share; eg.:
  - ethical concerns
  - commercial concerns
- Step 2: Choose a **data repository** or other sharing platform
  - overview: [re3data](#), [NIH](#), [FAIRsharing](#)
  - examples: [Zenodo](#), [Dryad](#)
- Step 3: Choose a **licence** and link to your paper and code; e.g.:
  - [Creative Commons](#)
  - [Open Data Commons](#)
- Step 4: **Upload** your data and documentation
  - good file organisation
  - appropriate file formats (e.g. `csv` > `xlsx`)
- Excel issues
  - formatting
  - thousands separators: , vs .
  - date conversion
  - formulas

---

## Sharing social media data

---

### Obstacles to data-sharing

- Reason 1: Preparing data for sharing is **resource-intensive**
- Reason 2: Not enough **credit** for data sharing
- Reason 3: Lack of **confidence** and **knowledge**
- Reason 4: Data protection **laws**
- Reason 5: Platform **terms of service**
- Reason 6: **Copyright**
- Reason 7: Informed **consent**

## ORIGINAL RESEARCH article

Front. Big Data, 16 January 2023  
Sec. Data Mining and Management  
Volume 5 - 2022 | <https://doi.org/10.3389/fdata.2022.971974>

This article is part of the Research Topic  
Social Recommendations and Applications for Metaverse  
[View all 3 Articles >](#)

[Download Article ▾](#)

1,676 Total views    480 Downloads    1 Citations [\(i\)](#)

[View article impact >](#)

4 [View altmetric score >](#)

### Edited by

 Ingmar Weber  
Saarland University, Germany

### Reviewed by

 Katja Mayer  
University of Vienna, Austria

 Lucia Ratiu  
Babeş-Bolyai University, Romania

 Tamara Gajic  
Geographical Institute, Ljubljana University

# Sharing social media data: The role of past experiences, attitudes, norms, and perceived behavioral control

 Esra Akdeniz<sup>1\*</sup>  Kerrin Emilia Borschewski<sup>1</sup>  Johannes Breuer<sup>2,3</sup>  Yevhen Voronin<sup>1†</sup>

<sup>1</sup> Data Services for the Social Sciences, GESIS - Leibniz Institute for the Social Sciences, Cologne, Germany

<sup>2</sup> Survey Data Curation, GESIS - Leibniz Institute for the Social Sciences, Cologne, Germany

<sup>3</sup> Center for Advanced Internet Studies (CAIS), GESIS - Leibniz Institute for the Social Sciences, Cologne, Germany

Social media data (SMD) have become an important data source in the social sciences. The purpose of this paper is to investigate the experiences and practices of researchers working with SMD in their research and gain insights into researchers' sharing behavior and influencing factors for their decisions. To achieve these aims, we conducted a survey study among researchers working with SMD. The questionnaire covered different topics related to accessing, (re)using, and sharing SMD. To examine attitudes toward data sharing, perceived subjective norms, and

Figure 31: [source](#)

- Reason 8: **Ethical** challenges
- Reason 9: Lack of common **standards**

[source](#)

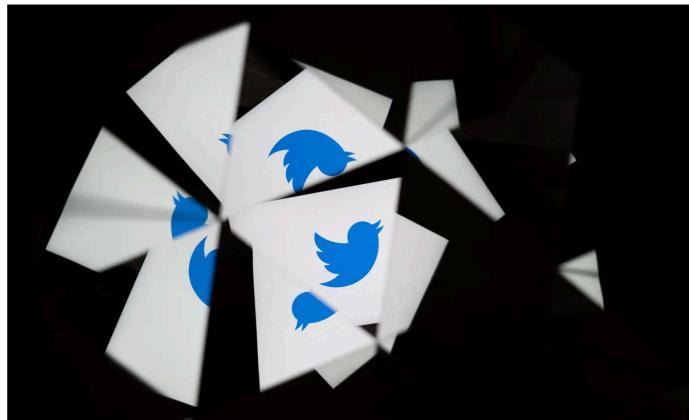
- additional reason: sharing not **considered** (necessary)
  - **ethical** and **privacy**: paper on incest on Twitter
- 

## The case of Twitter

- stage 1: access costly & legal grey area for scraping
- stage 2: Research API
- stage 3: Elon Musk → X → ...

SCIENCE / TWITTER - X / TECH

## Twitter just closed the book on academic research



/ Twitter was once an indispensable resource for academic research. That's changed under Elon Musk.

By [Justine Calma](#), a science reporter covering the environment, climate energy with a decade of experience. She is also the host of the Hell podcast.

May 31, 2023, 3:19 PM GMT+2 | □

Figure 32: [source](#)

---

## **Data and methods**

get active: see <https://tinyurl.com/opnsci>

---

---

## **Publishing**

### **Open access**

[The Turing Way tutorial on open access](#)

---

### **Routes to open access publishing**

---

## **Preregistration**

### **What is preregistration?**

When you preregister your research, you're simply specifying your **research plan in advance** of your study and submitting it to a registry.

Preregistration separates **hypothesis-generating** (exploratory) from **hypothesis-testing** (confirmatory) research.

- Both are important.
- But the same data cannot be used to generate and test a hypothesis, which can happen unintentionally and reduce the credibility of your results.
- Addressing this problem through planning improves the quality and transparency of your research.
- This helps you clearly report your study and helps others who may wish to build on it.

[Open Science Center tutorial](#)

---

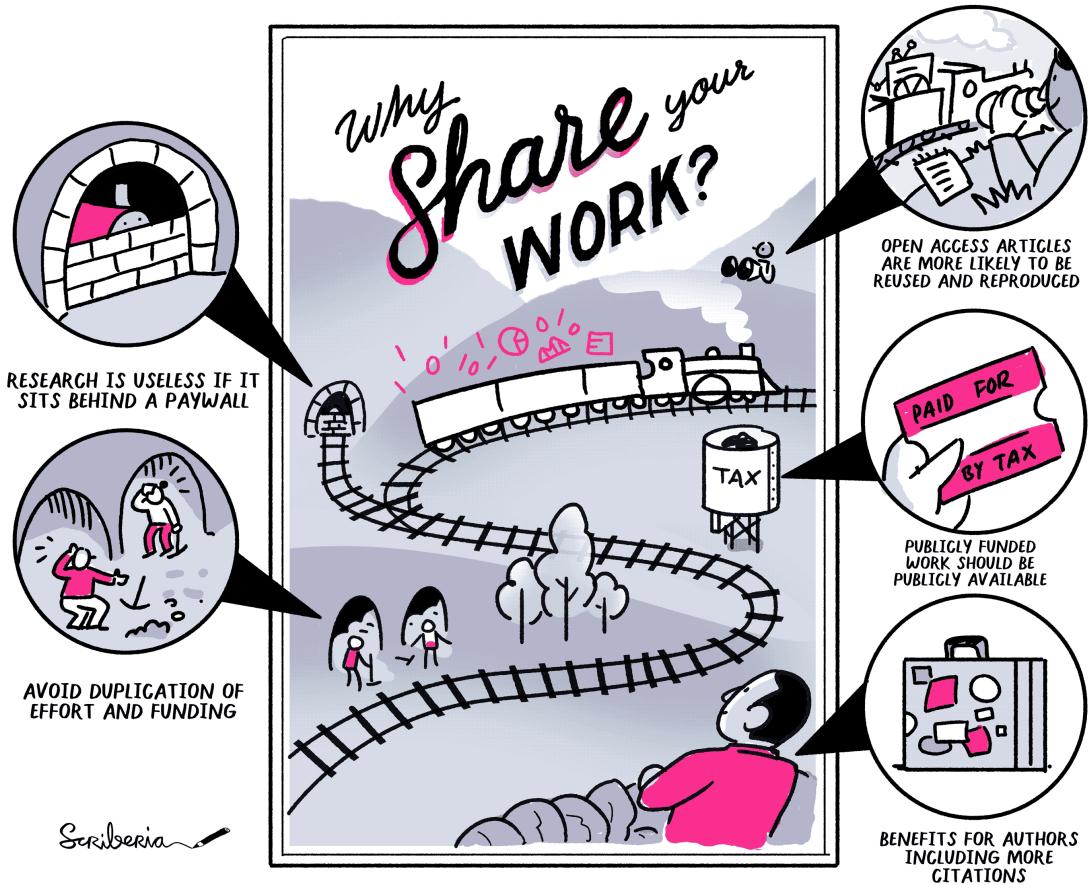


Figure 33: [source](#)

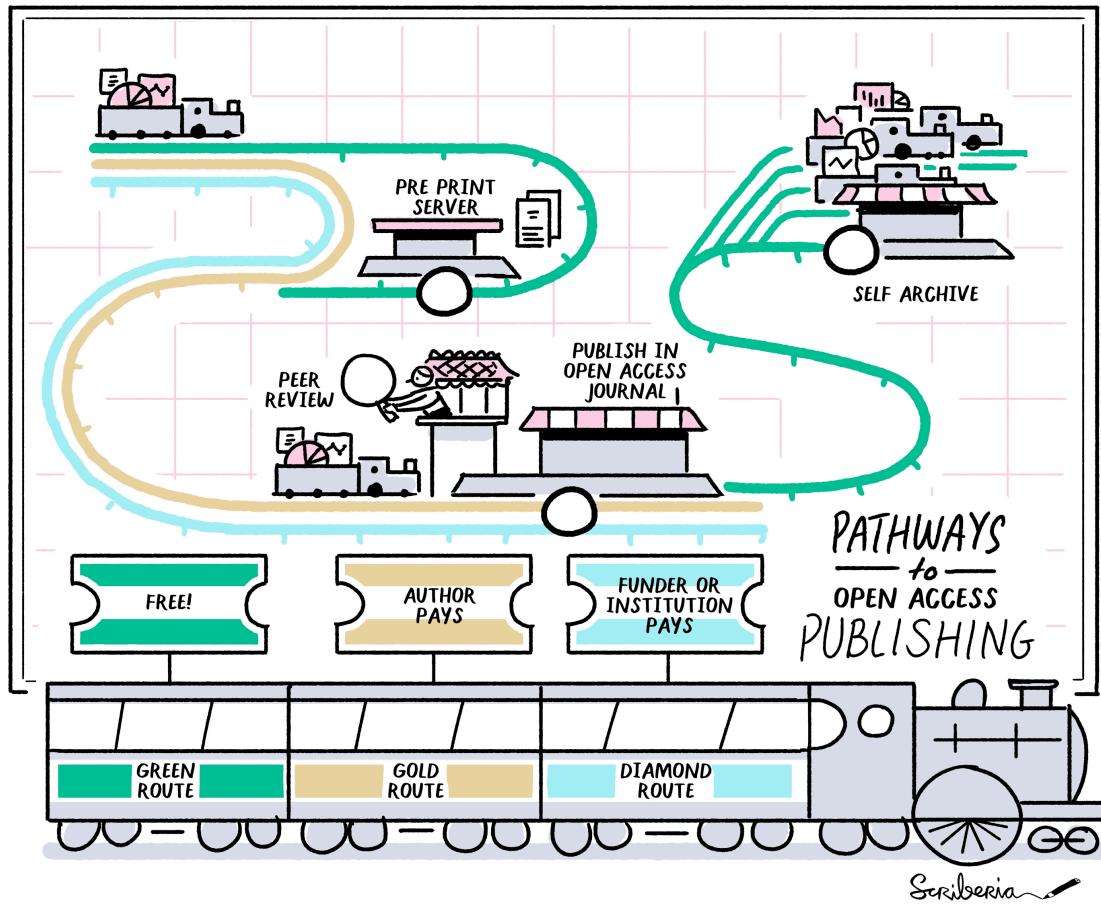


Figure 34: [source](#)

### The preregistration process



Figure 35: [source](#)

### Avoiding pitfalls through preregistration

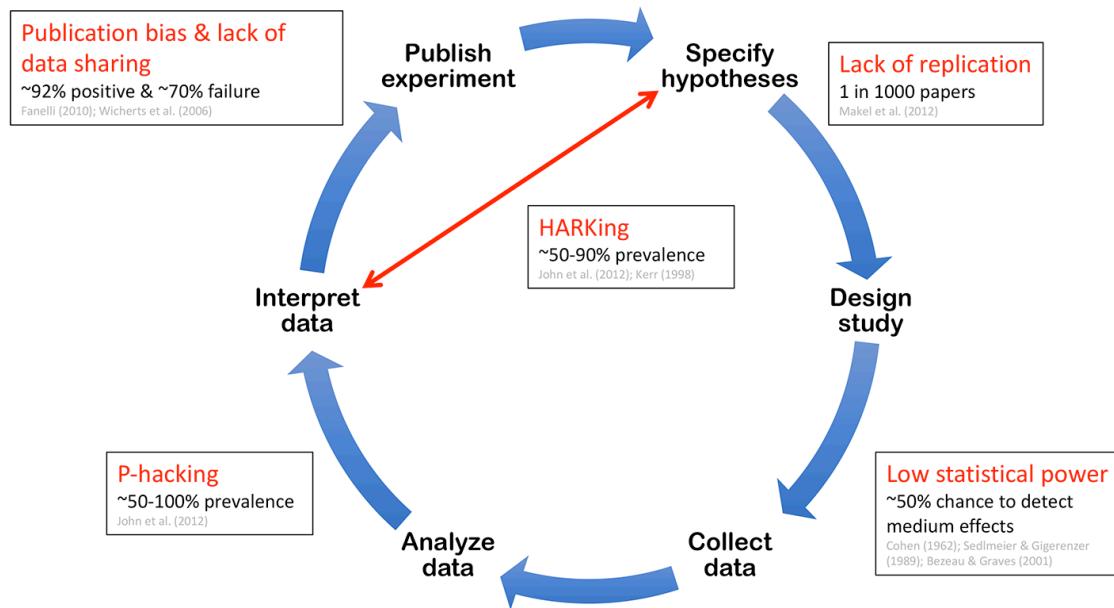


Figure 36: [source](#)

*HARKing*: hypothesizing after results are known

## **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

...

---

## **Authoring**

---

### **Plain text authoring**

---

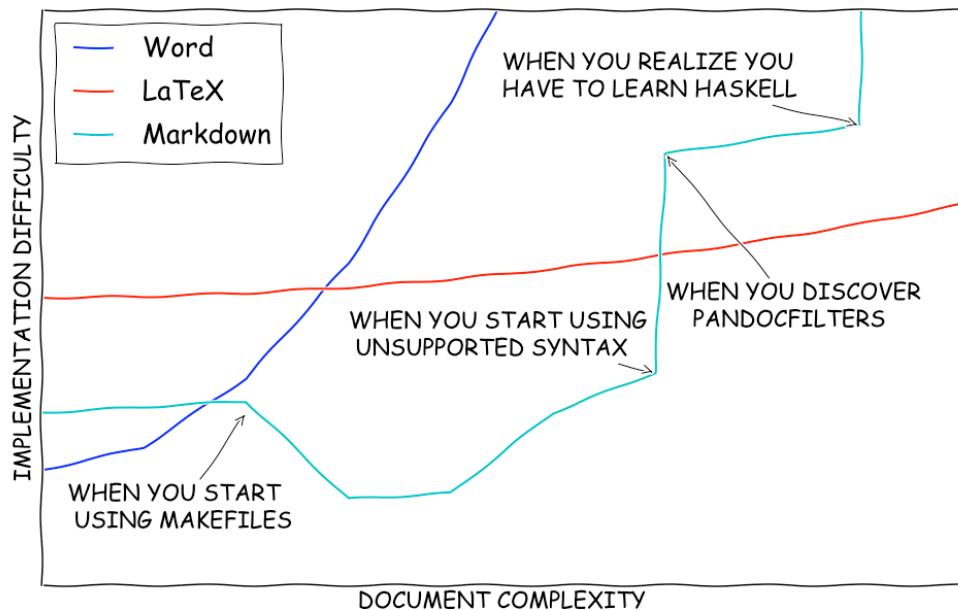


Figure 37: [source](#)

## Markdown syntax

# Text Formatting

Markdown Syntax	Output
<code>*italics*, **bold**, ***bold </code> <code>italics***</code>	<i>italics, bold, bold italics</i>
<code>superscript^2^ / subscript~2~</code>	<sup>superscript<sup>2</sup></sup> / <sub>subscript<sub>2</sub></sub>
<code>~~strikethrough~~ </code>	<del>strikethrough</del>
<code>`verbatim code` </code>	<code>verbatim code</code>

[Quarto: Markdown basics](#)

## Benefits of Markdown

- Markdown provides **semantic meaning** for content in a relatively simple way
- You can write rich formatted content extremely **quickly** (compared to writing directly in HTML tags)
- You can **read** Markdown easily in plain text before rendered by HTML
- It doesn't interrupt your **workflow** with the need to click buttons
- It's **platform-agnostic** so your content is not tied to the format of your editor

## Quarto

<https://quarto.org/>

# Welcome to Quarto®

## An open-source scientific and technical publishing system

- Author using [Jupyter](#) notebooks or with plain text markdown in your favorite editor.
- Create dynamic content with [Python](#), [R](#), [Julia](#), and [Observable](#).
- Publish reproducible, production quality articles, presentations, websites, blogs, and books in HTML, PDF, MS Word, ePub, and more.
- Share knowledge and insights organization-wide by publishing to [Posit Connect](#), [Confluence](#), or other publishing systems.
- Write using [Pandoc](#) markdown, including equations, citations, crossrefs, figure panels, callouts, advanced layout, and more.

## Multi-purpose publishing

### Dynamic Documents

Generate dynamic output using Python, R, Julia, and Observable. Create reproducible documents that can be regenerated when underlying assumptions or data change.

[Learn more »](#)

### Authoring Tools

Use your favorite tools including VS Code, RStudio, Jupyter Lab, or any text editor. Use the Quarto visual markdown editor for long-form documents.

[Learn more »](#)

### Beautiful Publications

Publish high-quality articles, reports, presentations, websites, and books in HTML, PDF, MS Word, ePub, and more. Use a single source document to target multiple formats.

[Learn more »](#)

### Interactivity

Engage readers by adding interactive data exploration to your documents using Jupyter Widgets, htmlwidgets for R, Observable JS, and Shiny.

[Learn more »](#)

### Scientific Markdown

Pandoc markdown has excellent support for LaTeX equations and citations. Quarto adds extensions for cross-references, figure panels, callouts, advanced page layout, and more.

[Learn more »](#)

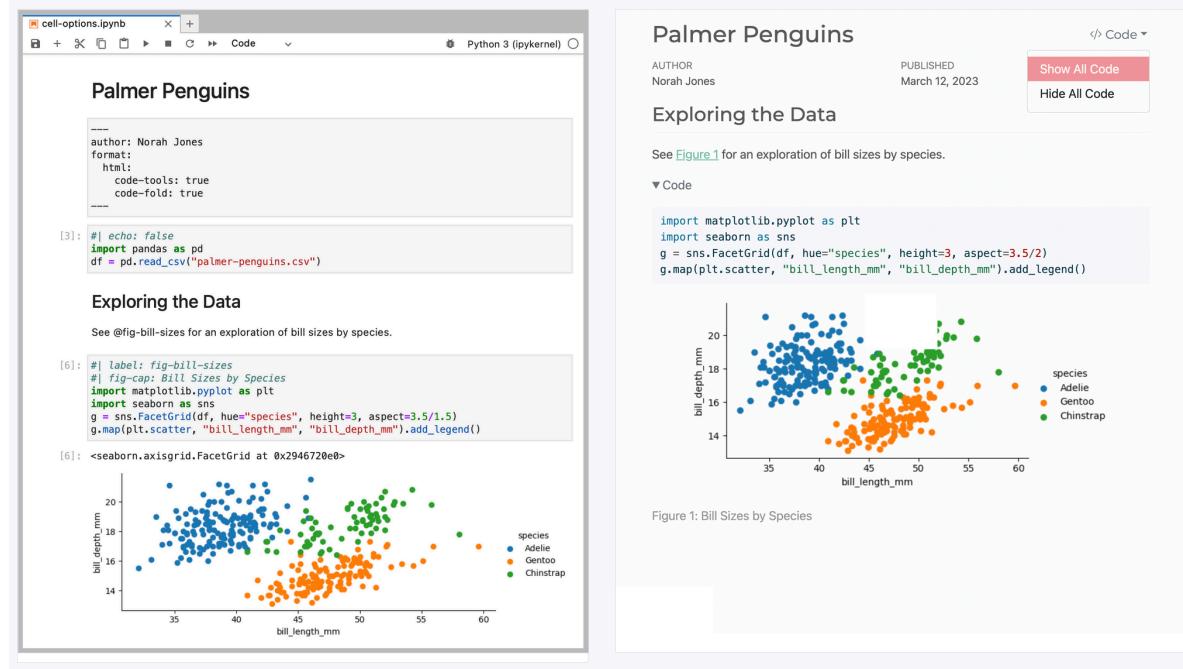
### Websites and Books

Publish collections of documents as a blog or full website. Create books and manuscripts in both print formats (PDF and MS Word) and online formats (HTML and ePub).

[Learn more »](#)

## Code

### Python



## R

```

---
title: "ggplot2 demo"
author: "Norah Jones"
date: "5/22/2021"
format:
  html:
    fig-width: 8
    fig-height: 4
    code-fold: true
---

## Air Quality

@fig-airquality further explores the impact of
temperature on ozone level.

```{r}
#| label: fig-airquality
#| fig-cap: "Temperature and ozone level."
#| warning: false

library(ggplot2)

ggplot(airquality, aes(Temp, Ozone)) +
  geom_point() +
  geom_smooth(method = "loess")
```
```

```

## ggplot2 demo

Norah Jones  
May 22nd, 2021

### Air Quality

[Figure 1](#) further explores the impact of temperature on ozone level.

► Code

Figure 1: Temperature and ozone level.

## Multi-format publishing

see the [Quarto gallery](#)

## Articles

[HTML for web publishing](#)

Interactivity

You can also add interactive plots. For example:

```
library(ggplot2)
ggplot(airquality, aes(Temp, Ozone)) +
  geom_point() +
  geom_smooth(method = "loess")
```

Figure 2: New Haven Temperatures

[PDF for high quality print](#)

Your Logo

The SocioEconomic Aspects of Stock Assessments

Recommendations for Increasing Assessment Accuracy and Improving Management Advice

Jane Doe<sup>1,2</sup>, Eva Novaković<sup>3</sup>, Matti Mekäläinen<sup>4,\*</sup> and Adhok Kumar<sup>2,3</sup>

1. Minnesota Department of Natural Resources, 500 Lafayette Road Saint Paul, MN 55155  
2. University of Minnesota, Department of Mathematics  
3. College of Law, University of Minnesota, Minneapolis, Minnesota, United States  
4. University of Kuopio, Department of Biological and Environmental Science, Kyröläntie 79, 90120, KEMIJÄRVI, Finland

[MS Word for Office workflows](#)

```
import numpy as np
import matplotlib.pyplot as plt

r = np.arange(0, 2, 0.01)
theta = 2 * np.pi * r
fig, ax = plt.subplots(subplot_kw={'projection': 'polar'})
ax.plot(theta, r)
ax.set_rticks([0.5, 1, 1.5, 2])
ax.grid(True)
plt.show()
```

Figure 1: A line plot on a polar axis

## Presentations

The image displays three presentation slides side-by-side, each representing a different presentation format:

- RevealJS HTML presentations:** A slide titled "Quarto Presentations" with the subtitle "Create beautiful interactive slide decks with Reveal.js". It features a dark background with white text and a small navigation bar at the bottom.
- Beamer PDF presentations:** A slide titled "Pop Songs and Political Science" by Steven V. Miller from the Department of Political Science at Clemson University. It includes the university's logo and a navigation bar at the bottom.
- PowerPoint presentations:** A slide titled "Best Practices for Administering RStudio in Production" by Nathan Stephens. It has a blue vertical bar on the left and a light blue background on the right, with a navigation bar at the bottom.

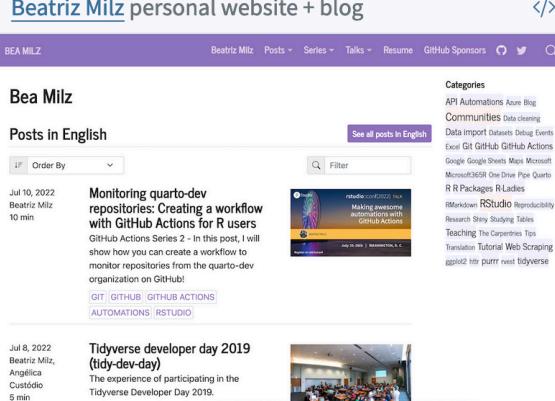
## Websites

The image shows two website interfaces:

- NASA & OpenScapes website:** A white-themed website for NASA OpenScapes. It features the NASA logo and the OpenScapes logo, followed by the text "NASA Openscapes". Below this, there is a paragraph about Earth science and data migration, links to project announcements and blog posts, and a call to connect on Twitter.
- nbdev from fast.ai:** A dark-themed website for nbdev. It features a large header with the text "Create delightful software with Jupyter Notebooks". Below the header, there is a "Get started" button and a section showing screenshots of various software interfaces integrated with nbdev, including GitHub, Jupyter, and Quarto.

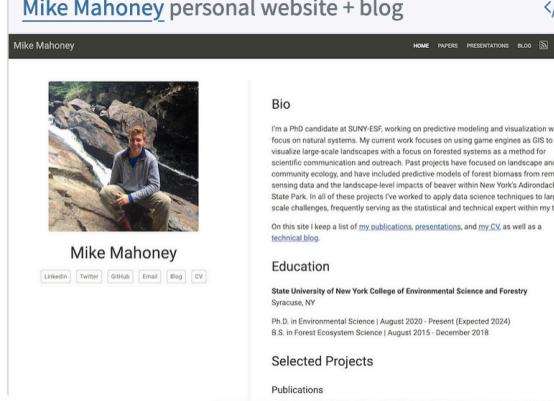
## Personal websites and blogs

[Beatriz Milz personal website + blog](#)



This screenshot shows a purple-themed blog interface. At the top, there's a navigation bar with links for 'Beatriz Milz', 'Posts', 'Series', 'Talks', 'Resume', and 'GitHub Sponsors'. Below the navigation, a sidebar on the left lists posts by date and category. A main content area displays a post titled 'Monitoring quarto-dev repositories: Creating a workflow with GitHub Actions for R users' with a preview image and a snippet of code.

[Mike Mahoney personal website + blog](#)



This screenshot shows a white-themed blog interface. At the top, there's a navigation bar with links for 'HOME', 'PAPERS', 'PRESENTATIONS', 'BLOG', and a search icon. Below the navigation, a sidebar on the left lists posts by date and category. A main content area displays a bio for Mike Mahoney, mentioning his PhD work at SUNY-ESF and his role as a PhD candidate. It also includes sections for 'Education' (State University of New York College of Environmental Science and Forestry, Syracuse, NY) and 'Selected Projects'.

## Demo

Demo repository: <https://github.com/wuqui/opensciwsdemo>

## CAIS workshop demo

AUTHOR  
qw

On this page

[Heading 1](#)  
[Code](#)  
[Lists](#)  
[Heading 2](#)

## Heading 1

### Code

```
for i in range(23):  
    print('servus')
```

### Lists

bullets

- one
- two
- three

---

## **Authoring a Quarto document**

You have until 16:00 h to,

based on Quarto's excellent [getting started guide](#),

- install Quarto
- make an example text document using Quarto Markdown syntax (optional: clone [my demo repo](#))
- `render` the document to multiple output formats (`pdf`, `html`, `docx`, etc.)

further steps

- publishing your document(s) as a website on GitHub pages (see [tutorial](#))
  - refactoring existing document (e.g. paper, website) to Quarto Markdown
- 

## **Recap, open issues, feedback**

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