

How to make research reproducible?

Agata Bochynska, PhD

Open Research and Digital Scholarship Center
University of Oslo Library

@AgataBochynska

agata.bochynska@ub.uio.no



Time and place: Mar. 7, 2024 10:00 AM – 11:00 AM, Zoom

Open and reproducible research: An overview

Learn about what open research is and how to make your own research more transparent and reproducible.



Time and place: Mar. 8, 2024 10:00 AM – 12:00 PM, Zoom

How to preregister research studies?

Learn about what preregistration is and how to preregister your own studies.



Time and place: Mar. 11, 2024 10:00 AM – 11:00 AM, Zoom

How to make research reproducible?

Learn about tools and practices for more reproducible and effective research.



Time and place: Mar. 14, 2024 10:00 AM – 11:30 AM, Zoom

How to publish openly?

Learn about preprints, peer-review process, Open Access and how can you choose the best way to publish your results openly.



Time and place: Mar. 15, 2024 10:00 AM – 11:30 AM, Zoom

How to make research more visible?

Learn about different tools, platforms and services to share your research and other contributions, and how you utilise them to make yourself and your work more visible to the academic community and the society at large.

Open and reproducible research courses

March 7th – 15th

Roadmap

- Some definitions
- Open research and reproducibility
- Reproducible data acquisition, processing, analyses and reports/publications (with some useful tools)
- Take-aways
- Q&A time!

		DATA	
		Same	Different
ANALYSIS	Same	Reproduced	Replicated
	Different	Robust	Generalized

Reproduced

results are consistent when following the same method and analysis steps with the **same input data**

Quantitative studies

computational reproducibility

Qualitative studies

process transparency

Quantitative studies

computational reproducibility

“obtaining consistent computational results using the same input data, computational steps, methods, code, and conditions of analysis”

Quantitative studies

computational reproducibility

Re-running analyses/code with the same data

Qualitative studies

process transparency

Arriving at similar (consistent) interpretation by following the same analysis process

Qualitative studies

process transparency

Following the step-by-step reasoning and interpretation process

Reproducibility is strongly
associated with **transparency**

“**Open Science** has the potential of making the scientific process more **transparent, inclusive** and **democratic**. It is (...) a true game changer in bridging the science, technology and innovation gaps and fulfilling the **human right to science**.”

https://youtu.be/I3Wkvx_ZaFo

<https://www.unesco.org/en/natural-sciences/open-science>



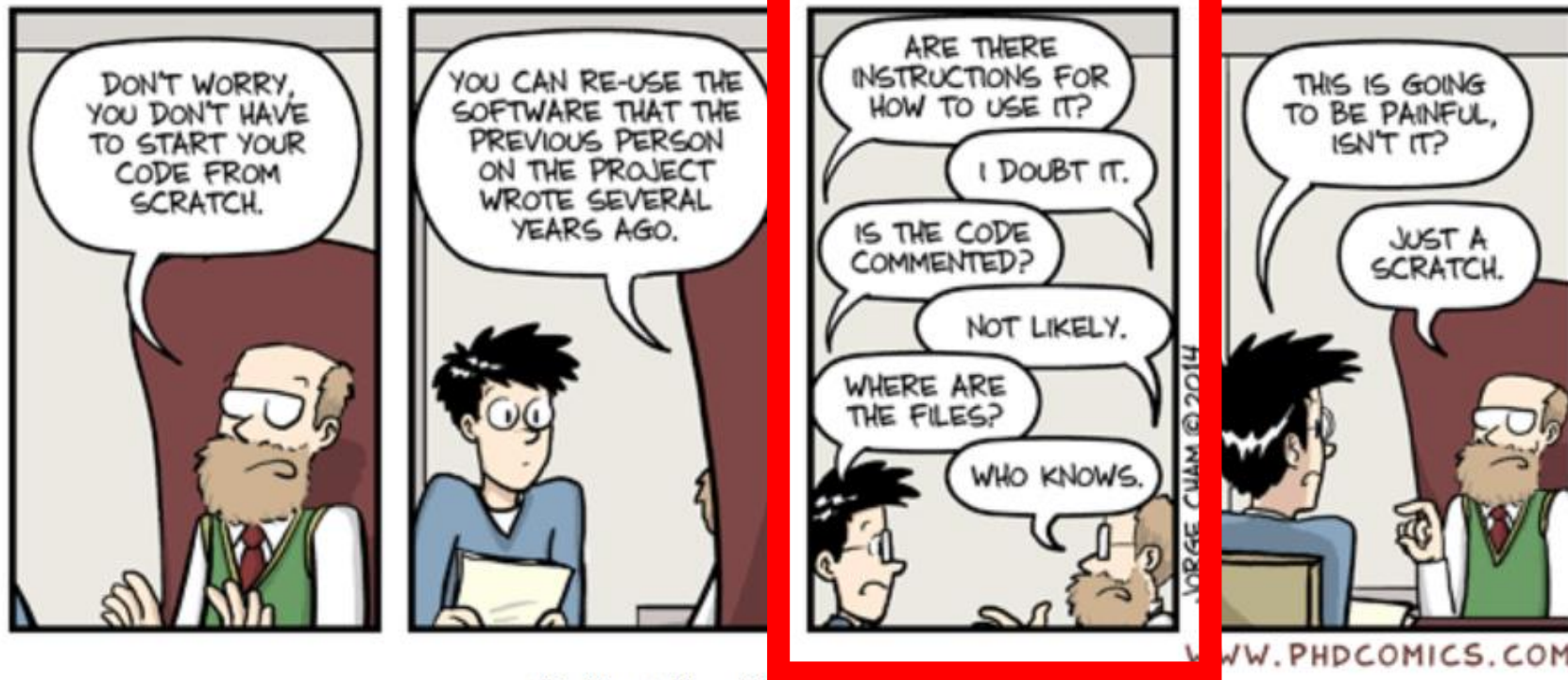
**UNESCO Recommendation
on Open Science**

Reproducible does not
(have to) mean fully **open**

As open as possible,
as closed as necessary

Open but not usable

Piled Higher and Deeper by Jorge Cham www.phdcomics.com



title: "Scratch" - originally published 3/12/2014

Reasons for irreproducibility:

- Unavailability of materials, data and/or analyses
- Poor data management
- Unclear analysis specification
- Lack of documentation
- Errors in reporting numbers
- Lack of quality checking procedures
- Insufficient peer review

Reproducible research workflows

Data acquisition and processing

Data analyses

Data reports (manuscripts)

Data acquisition and processing

Data organization

Data documentation

Version control

Organized data

```
project_name/
├── README.md           # overview of the project
├── data/              # data files used in the project
│   ├── README.md     # describes where data came from
│   └── sub-folder/   # may contain subdirectories
├── processed_data/   # intermediate files from the analysis
├── manuscript/       # manuscript describing the results
├── results/          # results of the analysis (data, tables, figures)
├── src/              # contains all code in the project
│   ├── LICENSE       # license for your code
│   ├── requirements.txt # software requirements and dependencies
│   └── ...
└── doc/              # documentation for your project
    ├── index.rst
    └── ...
```

Research project with a proper data file structure. Image taken from CodeRefinery, Lesson on Reproducible Research. Shared under CC-BY 4.0.

Versioned data

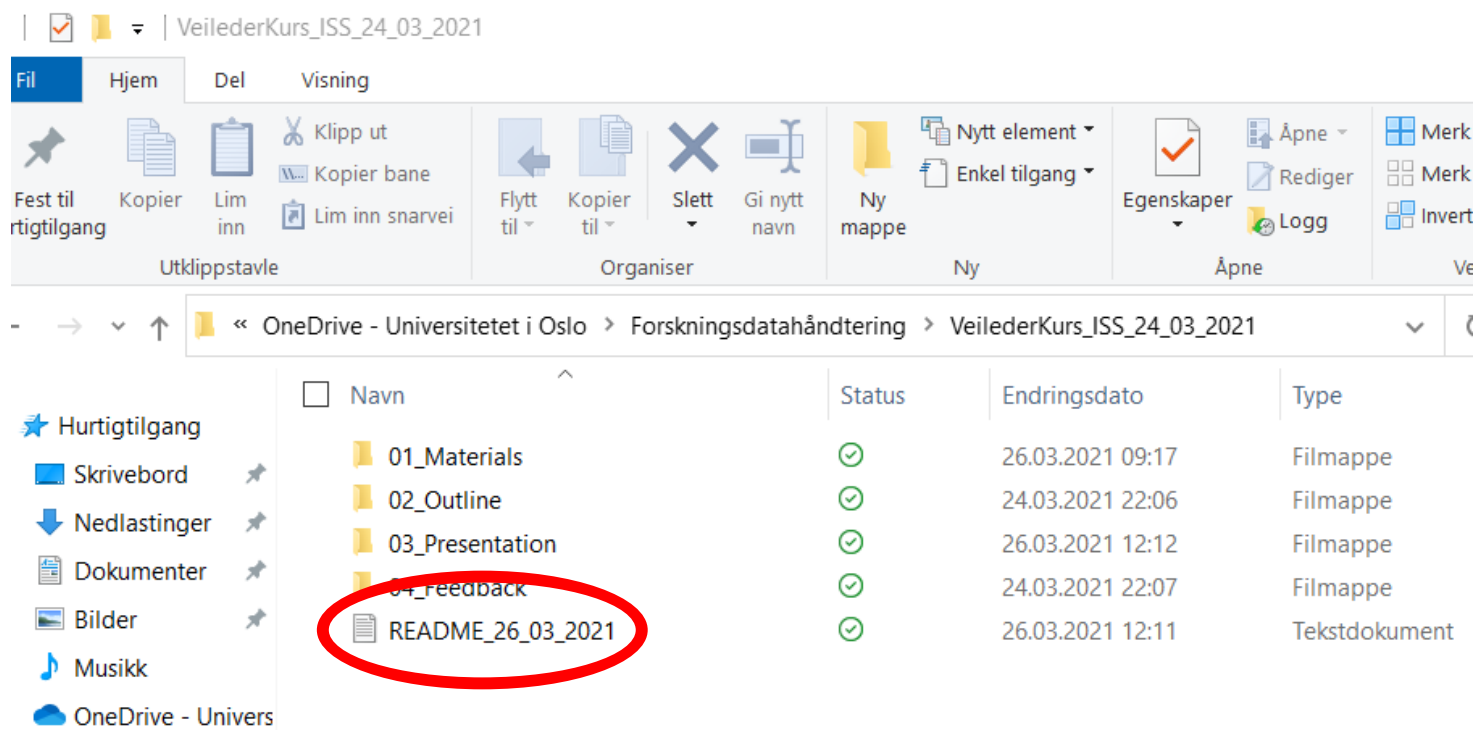
Versioning refers to saving **new copies** of your files when you make changes so that later you can go back and **retrieve** specific **versions** of your files

- DataFileName_1.0 = original document
- DataFileName_1.1 = original document with minor revisions
- DataFileName_2.0 = document with substantial revisions



Documented data: README-files

- The first file to open
- Map for navigating and exploring files and their content
- One README.txt file per folder



Documented data: CODEBOOK

- Explains all variables and their codes in the dataset
- It typically contains:
 - variable names, variable labels, variable codes, variable formats, missing data (in quantitative research)
 - codes, code definitions, examples of what to include with a given code (in qualitative research)
- Can be also called **Data Dictionary**

Tools that help: Templates

- Cornell University template and guide to README.txt-files:

<https://data.research.cornell.edu/content/readme>

- README.txt-files from DataverseNO:

[General template](#)

[Example for social sciences](#)

[Example for life sciences](#)

Tools that help: Nettskjema codebook

[View](#) [Form builder](#) **[Codebook](#)** [Settings](#) [Collect responses](#) [See results](#)

Codebook

Mapping questions and alternatives to variables is necessary if the results of a survey are going to be processed in an external analysis tool (e.g. SPSS, STATA or R).

[Read more about, and get an introduction to the codebook in Nettskjema](#)

 [Download codebook as text](#)

 [Download codebook as SPSS syntax file](#)

Tools that help: ELN

Electronic Lab Notebooks

help document research, experiments and procedures
performed in laboratories



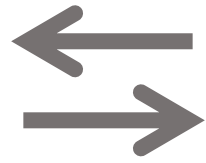
Data analysis

Step-by-step documentation

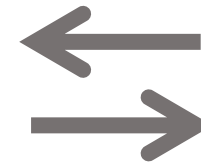
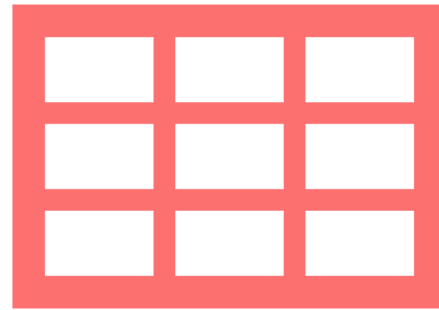
Version control

Cloud computing and/or containers

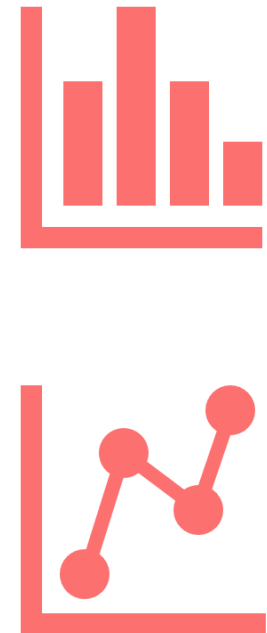
Raw data



Processed data



Analyzed data



Qualitative studies process transparency

Following the step-by-step reasoning and
interpretation process

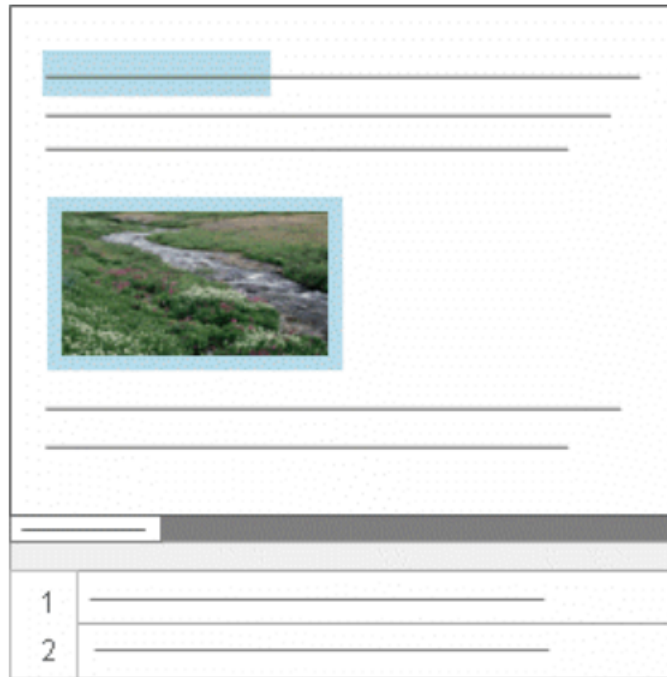
Tools that help: Annotations



Use annotations to comment on selected parts of a source or node

Like scribbled notes in the margin, annotations let you record comments, reminders or observations about specific content in a source or node.

Annotated content is highlighted in blue and the text of the annotation is displayed in the **Annotations** tab at the bottom of the window.



Annotation for Transparent Inquiry (ATI) at a Glance

Annotation for Transparent Inquiry (ATI)

ATI Models

ATI Instructions

Why ATI?

Empowering Openness in Law-Related Research: A Pilot

Working with Sensitive Research Data (WSRD)

Publications

Annotation for Transparent Inquiry (ATI) facilitates transparency in qualitative research by allowing scholars to “annotate” specific passages in an article. Annotations amplify the text and, when possible, include a link to one or more data sources underlying a claim; data sources are housed in a repository.

Any digitally published manuscript can be annotated using ATI (here: an article in *International Organization* published by Cambridge University Press)

Hungary, this was their only stated concern. However, many states conditioned their recognition decision on an action related to Indian troop withdrawal and gave three different types of reasons for doing so. States also differed in the extent of troop withdrawal they required before recognition. See [Table 2](#) for a full list of states, their stated reason for conditioning recognition on withdrawal (if any can be identified), and what recognition was conditioned on (whether actual withdrawal or a proxy).

The first type of reason, opposition to condoning or legitimizing aggression, is labeled as “Non-aggression.” A good example comes from Mexican Foreign Minister Emilio Óscar Rabasa who reported that the Mexican president had decided not to recognize Bangladesh because, “since the Mexicans, like many Latin Americans, refuse to condone territorial aggrandizement as a result of war, they would prefer to wait on the withdrawal of Indian troops as the sign of true independence.”⁸⁸

This statement also appeals to “true independence.” Self-determination is another important value expressed by the Mexican representative and is the second type of reason commonly appealed to as justifying recognition as Bangladesh. For

88. See [Figure 2](#).
 89. A frequent concern was that states had to recognize in a group, or on the same day as multiple other states. However, even allowing for minor coordination problems, this in and of itself cannot explain the length of time taken to make recognition decisions and declarations.
 90. Cable from Hope, 16 January 1972, FCO 37/1020.

The screenshot shows a digital manuscript with a highlighted passage: "they would prefer to wait on the withdrawal of Indian troops as the sign of true independence." An ATI annotation box is overlaid on this passage. The annotation contains the following information:

- QDR** CambridgeCore/ATI
- Annotation for Transparent Inquiry (ATI)**
- Full Citation:** Sir Peter Hope, UK Ambassador to Mexico, a confidential telegram from Hope to the Foreign and Commonwealth Office, 26 January 1972. Folder 37/1020 of the FCO Archives held at the National Archives at Kew, UK.
- Source Excerpt:** Rabasa said that, since the Mexicans, like many Latin Americans, refuse to condone territorial aggrandizement as a result of war, they would prefer to wait on the withdrawal of Indian troops as the sign of true independence.
- Analytic Note:** This is a confidential telegram from UK Ambassador to Mexico Sir Peter Hope to the Foreign and Commonwealth Office of 26 January, 1972, from folder 37/1020 of the FCO Archives held at the National Archives at Kew, UK. This excerpt shows that the Mexican Foreign Minister, Emilio Oscar Rabasa, gave as a reason for the Mexican President's decision not to recognize Bangladesh, that they did not want to condone territorial aggrandizement as a result of war until Indian troops had been withdrawn. The telegram also indicates that this reason and another reason, i.e. that Mujib's assumption of several cabinet portfolios cast doubt on the fact that his government had been elected by the people, were the only two reasons cited by the Mexican government.
- Data Source:** <https://data.beta.qdr.org/api/access/datafile/25297?key=13e4c93f-1172-4d53-8a07-6f2651e5da97>

ATI Annotation: Displayed alongside article. Created by author, curated by QDR, hosted and served by Hypothesis, displayed on publisher's web site

Elements of an ATI annotation: One or more of the following:

- Analytic Note
- Source Excerpt
- Source Excerpt Translation
- Link to Data Source

Link to data source housed in QDR

Any passage in the text or in notes of a manuscript can be annotated using ATI

SYMPOSIUM

Active Citation: A Precondition for Replicable Qualitative Research

Andrew Moravcsik, *Princeton University*

Quantitative studies

computational reproducibility

Re-running analyses/code with the same data

Tools that help: analysis via code



File Home Insert Draw Page Layout Formulas **Data** Review View Help

Get Data Refresh All Queries & Connections Properties Edit Links Stocks (En... Geography... Sort Filter Sort & Filter Clear Reapply Advanced Text to Columns Data Tools What-If Analysis Forecast Sheet Outline

Get & Transform Data Queries & Connections Data Types Sort & Filter Data Tools Forecast Outline

A1 1

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	1	A	0,657792															
2	2	A	0,443001															
3	3	A	0,490799															
4	4	A	0,315724															
5	5	A	0,86144															
6	6	A	0,481306															
7	7	A	0,337931															
8	8	A	0,126269															
9	9	A	0,571799															
10	10	A	0,317623															
11	11	B	0,285825															
12	12	B	0,931719															
13	13	B	0,615969															
14	14	B	0,068008															
15	15	B	0,918656															
16	16	B	0,531989															
17	17	B	0,695157															
18	18	B	0,810593															
19	19	B	0,989223															
20	20	B	0,54528															
21	21	B	0,4301															
22	22	B	0,130967															
23	23	B	0,384371															

```
57 {r}
58 # Main Analysis Data
59
60 #load wide format data and preview
61 sum_data <- read.csv("Data/Experiment2_SumData.csv")
62 head(sum_data)
63
64 #check summary statistics for the dataset
65 describe(sum_data)
66 {r}
```

```
67
68 {r}
69 # Main Analysis
70
71 #ttest on the total proportion looking to shape change against chance (0.5)
72 t.test(sum_data$ShapeProportion, mu=0.5)
73 sd(sum_data$ShapeProportion)
74 se <- sd(sum_data$ShapeProportion)/sqrt(length(sum_data$ShapeProportion))
75 se
76
77 #compute the effect size (Cohen's D)
78 cohensD(sum_data$ShapeProportion, mu=0.5)
79
80
81 # Bayesian ttest on the total proportion looking to shape change against chance (0.5)
82 testMain <- ttestBF(sum_data$ShapeProportion, mu=0.5)
83 testMain
84 sd(sum_data$ShapeProportion)
85 se <- sd(sum_data$ShapeProportion)/sqrt(length(sum_data$ShapeProportion))
86 se
87 {r}
```

Tools that help: version control

- Git: Free and open source version control system



- GitHub: is an internet hosting service for software development and version control using Git



<https://git-scm.com/>

<https://github.com/>

<https://youtu.be/gY2JwRfin1M>

Tools that help: shared notebooks



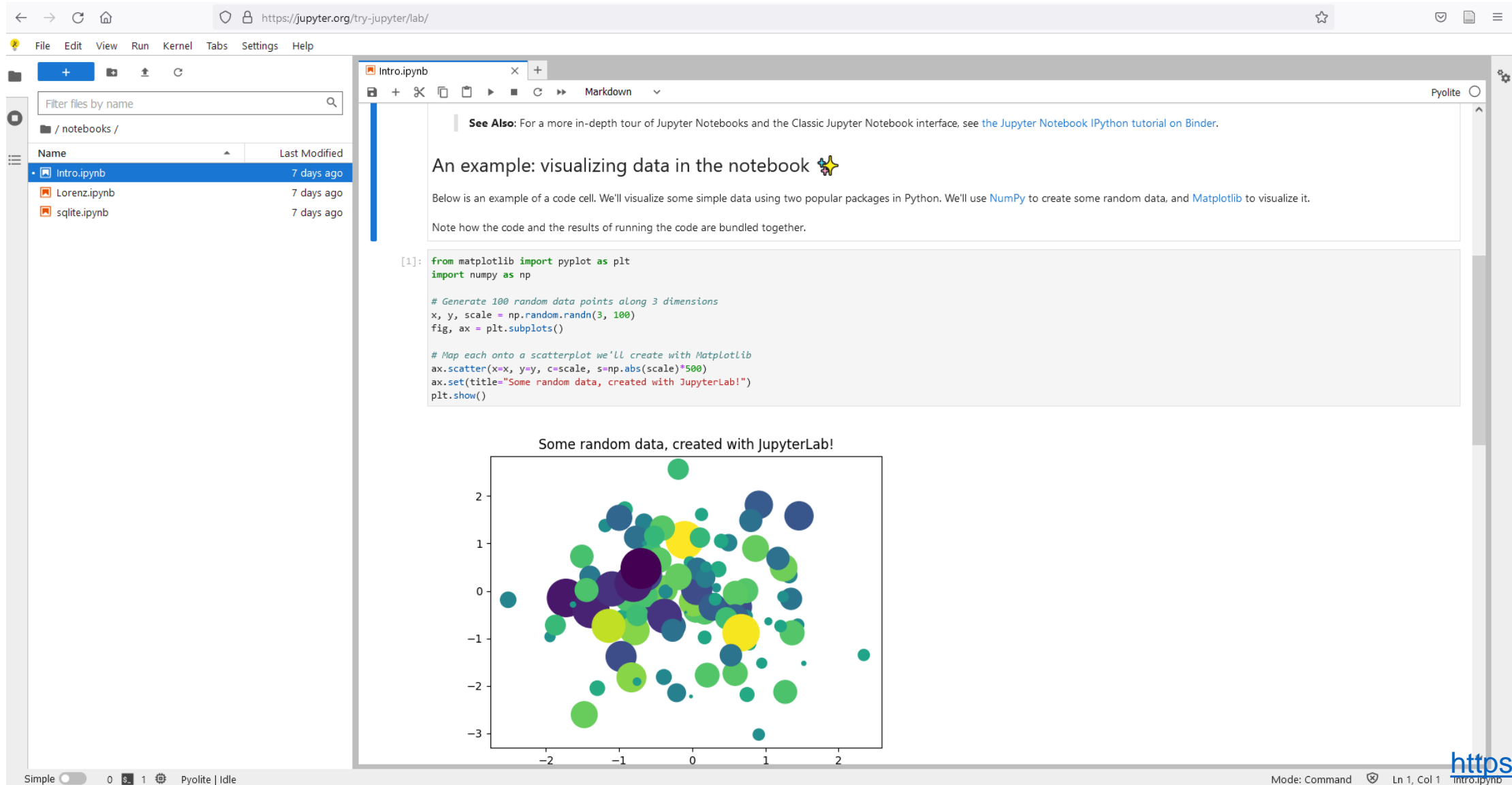
[Try](#) [Install](#) [Get Involved](#) [Documentation](#) [News](#) [Governance](#) [Security](#) [About](#)



Free software, open standards, and web services for interactive computing across all programming languages

<https://jupyter.org/>

Tools that help: shared notebooks



The screenshot displays the JupyterLab web interface. The browser address bar shows <https://jupyter.org/try-jupyter/lab/>. The left sidebar contains a file browser with a search bar and a list of notebooks: `intro.ipynb` (7 days ago), `Lorenz.ipynb` (7 days ago), and `sqlite.ipynb` (7 days ago). The main workspace shows a notebook titled `Intro.ipynb` in `Markdown` mode. It includes a `See Also` link, a heading `An example: visualizing data in the notebook`, and a code cell with the following Python code:

```
[1]: from matplotlib import pyplot as plt
import numpy as np

# Generate 100 random data points along 3 dimensions
x, y, scale = np.random.randn(3, 100)
fig, ax = plt.subplots()

# Map each onto a scatterplot we'll create with Matplotlib
ax.scatter(x=x, y=y, c=scale, s=np.abs(scale)*500)
ax.set(title="Some random data, created with JupyterLab!")
plt.show()
```

Below the code cell is a scatter plot titled "Some random data, created with JupyterLab!". The plot shows approximately 100 data points in a 2D space, where the x and y axes range from -3 to 2. The points are colored and sized based on a third variable, 'scale', resulting in a distribution of points with varying colors (including purple, blue, green, and yellow) and sizes.

At the bottom of the interface, the status bar shows `Simple` mode, `Pyolite | Idle`, and the current file `Intro.ipynb`. A URL <https://jupyter.org/> is visible in the bottom right corner.

Tools that help: Quality check

Language

Python

R

Shell/Bash

Static code analysis tool

[Pylint](#), [prospector](#)

[lintr](#)

[shellcheck](#)

Language

Python

R

Shell/Bash

HTML

Formatter Tool

[Black](#), [yapf](#)

[formatR](#)

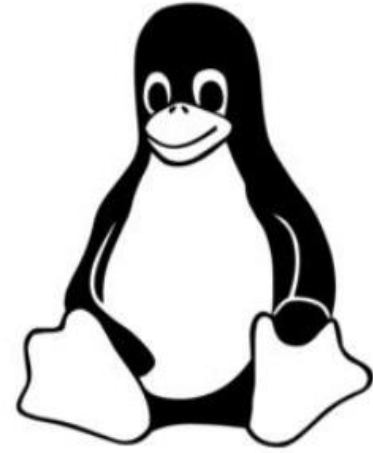
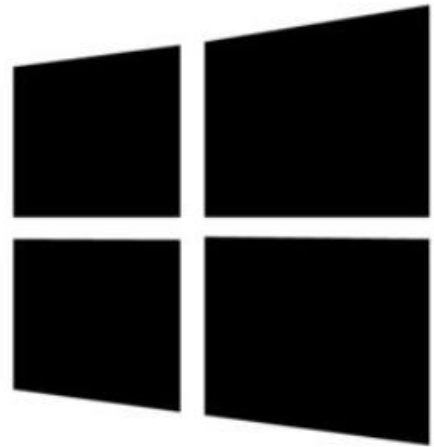
[ShellIndent](#)

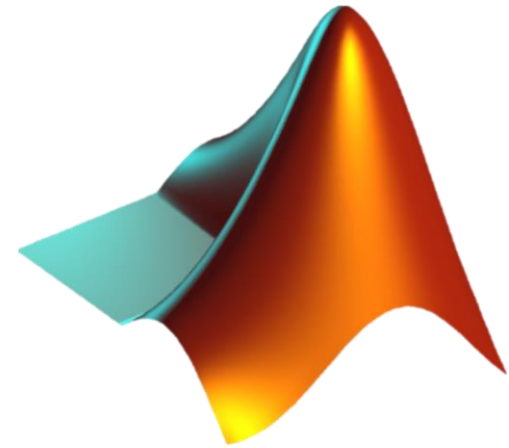
[Tidy](#)

Tools that help: Code review

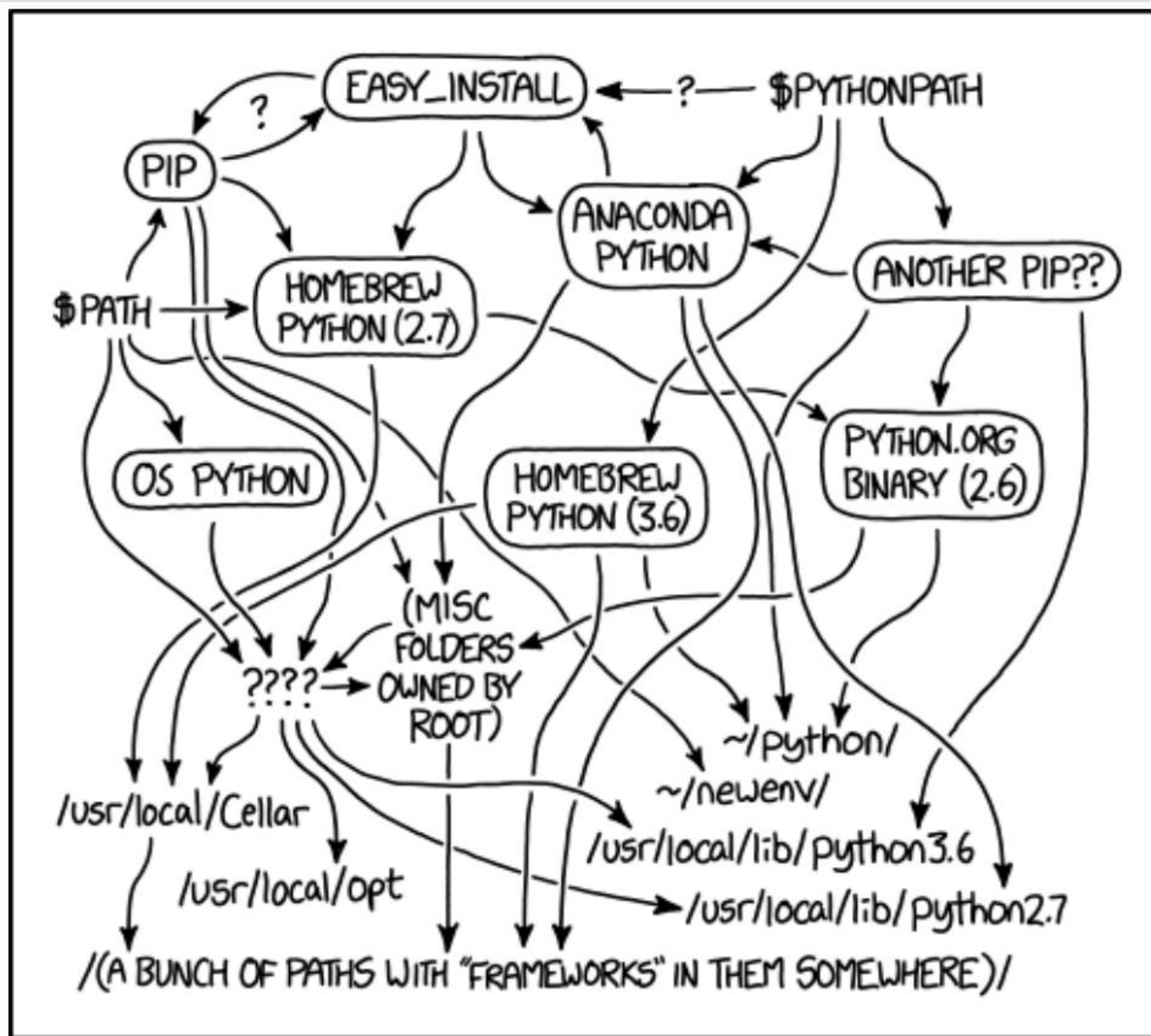


However...










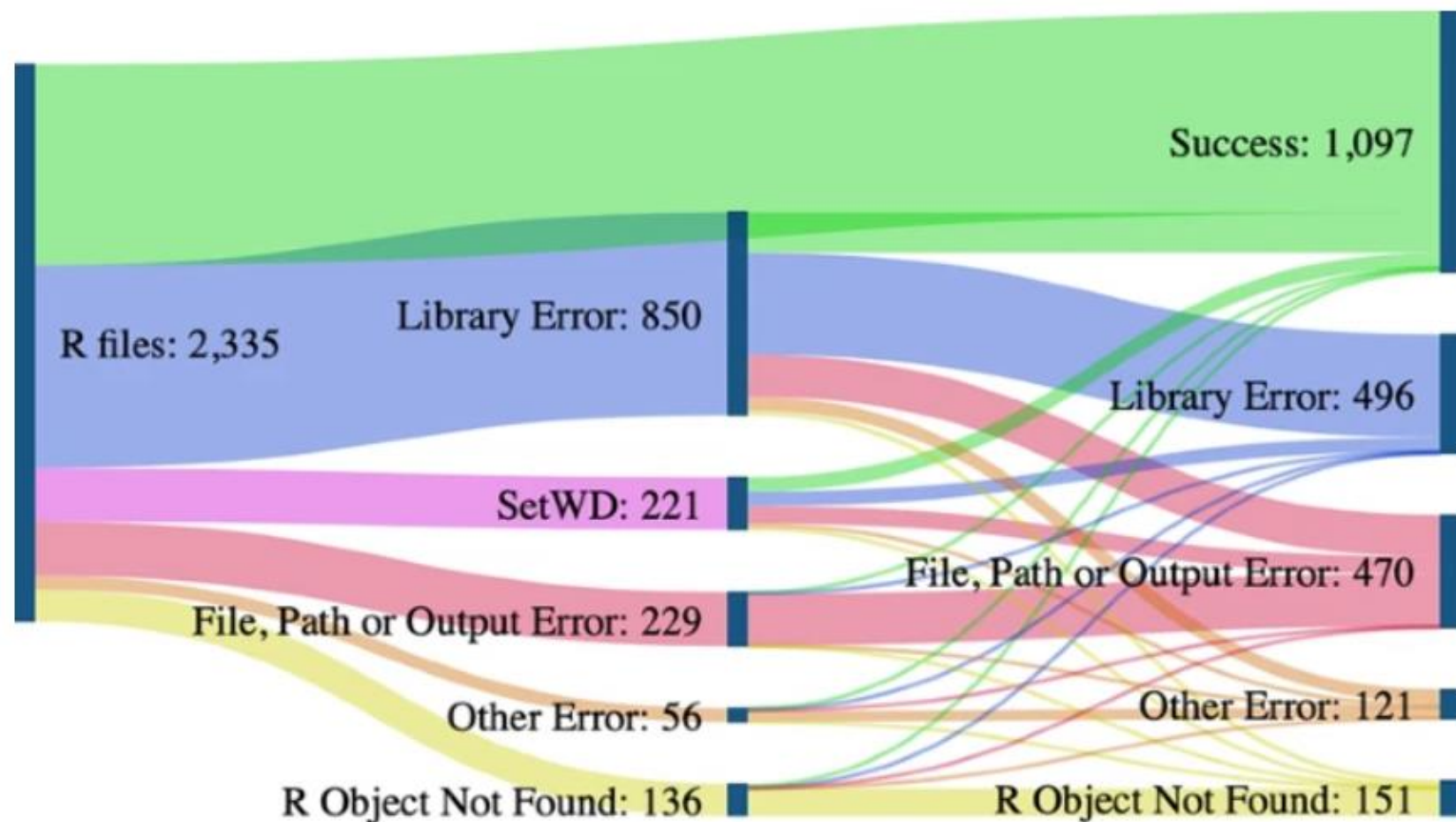
MY PYTHON ENVIRONMENT HAS BECOME SO DEGRADED THAT MY LAPTOP HAS BEEN DECLARED A SUPERFUND SITE.

A large-scale study on research code quality and execution

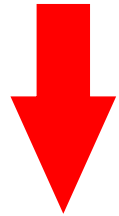
[Ana Trisovic](#) , [Matthew K. Lau](#), [Thomas Pasquier](#) & [Mercè Crosas](#)

[Scientific Data](#) **9**, Article number: 60 (2022) |

19k Accesses | **7** Citations | **399** Altmetric







Computational
environment



Container

Tools that help: Containers

Interaction style What is reproduced?	Graphical	Command line
Software & versions	 binder	 CONDA
Entire system		 docker

Private Untitled Capsule Oct 19, 2022 17:16

Capsule File Help

master Share

Environment x Metadata x Y: metadata.yml x Dockerfile x

Editing the Dockerfile will permanently disable the Environment Editor. Please commit before editing. [Use Environment Editor](#) [Unlock](#)

```

1 # hash:sha256:f137bdfbb7795cd46da870b7886b0ebcfe5d14a2070513da9f40f979e584cc81
2 FROM registry.codeocean.com/codeocean/r-studio:2022.07.0-548-r4.2.1-ubuntu18.04
3
4 ARG DEBIAN_FRONTEND=noninteractive
5
6 RUN Rscript -e 'remotes::install_version("ggplot2")' \
7     && Rscript -e 'remotes::install_version("tidyverse")'
8

```

Files

Core Files

- metadata 62 B
- Y: metadata.yml 62 B
- environment 195 B
- Dockerfile 195 B
- code 0 B
- data Manage Datasets 0 B

App Panel

Results

- results

Other Files

Upload

or

Start with Sample Files

Reproducible Run

or launch a cloud workstation

lab Studio jupyter Shiny

Timeline

Submit for publication...

What happens once I publish?

Select filter...

Agata Bochynska committed Oct 19, 2022

Added packages

Oct 19, 2022 Created Capsule



The Digital Lab for Computational Scientists

Start faster. Reproduce reliably. Focus on science.

<https://codeocean.com/>

Data reports (manuscripts)

Linked tables and analyses

Version control

Collaboration

Tools that help: R Markdown

R Markdown

from  Studio

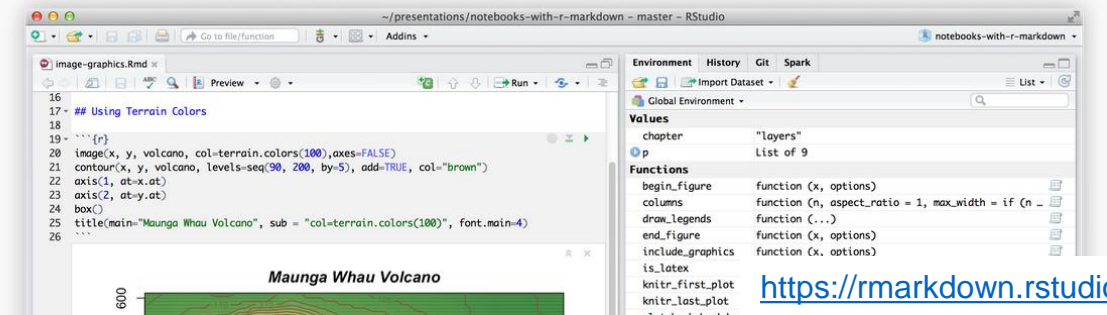
[Get Started](#) [Gallery](#) [Formats](#) [Articles](#) [Book](#) [References](#) 

Analyze. Share. Reproduce.

Your data tells a story. Tell it with R Markdown.

Turn your analyses into high quality documents, reports, presentations and dashboards.

R Markdown documents are fully reproducible. Use a productive [notebook interface](#) to weave together narrative text and code to produce



The screenshot shows the RStudio interface with a notebook. The code editor contains the following R code:

```
16  
17 # Using Terrain Colors  
18  
19 ...[r]  
20 image(x, y, volcano, col=terrain.colors(100), axes=FALSE)  
21 contour(x, y, volcano, levels=seq(30, 200, by=5), add=TRUE, col="brown")  
22 axis(1, at=x.at)  
23 axis(2, at=y.at)  
24 box()  
25 title(main="Maunga Whau Volcano", sub = "col=terrain.colors(100)", font.main=4)  
26 ...
```

The plot area shows a topographic map of Maunga Whau Volcano with contour lines. The right-hand pane displays the Environment and Functions windows. The Environment window shows the following values:

Object	Value
chapter	"Layers"
p	List of 9

The Functions window lists the following functions:

Function	Definition
begin_figure	function(x, options)
columns	function(n, aspect_ratio = 1, max_width = if (n ...
draw_legends	function(...)
end_figure	function(x, options)
include_graphics	function(x, options)
is_latex	
knitr_first_plot	
knitr_last_plot	

The URL <https://rmarkdown.rstudio.com/> is visible in the bottom right corner.

Tools that help: Quarto

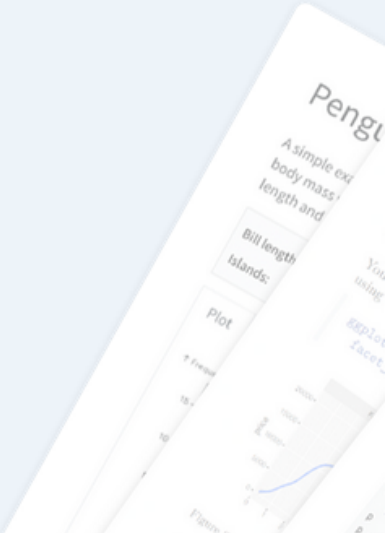


Overview Get Started Guide Extensions Reference Gallery Blog Help ▾

Welcome to Quarto

Quarto® is an open-source scientific and technical publishing system built on [Pandoc](#)

- Create dynamic content with [Python](#), [R](#), [Julia](#), and [Observable](#).
- Author documents as plain text markdown or [Jupyter](#) notebooks.
- Publish high-quality articles, reports, presentations, websites, blogs, and books in HTML, PDF, MS Word, ePub, and more.
- Author with scientific markdown, including equations, citations, crossrefs, figure panels, callouts, advanced layout, and more.



Other tools

- [Overleaf](#) (collaborative LaTeX editor)
- [HackMD](#) (a realtime web-based collaborative Markdown editor)
- [Manuscripts.io](#) (a collaborative authoring tool that support scientific content and reproducibility)
- [Rrtools](#) (instructions, templates, and functions for making a basic compendium suitable for writing a reproducible journal article or report with R)
- [Jupyter Notebooks](#) (can be used for supplementary material with journal articles).

Reproducible research workflows

Data acquisition and processing

Data analyses

Data reports (manuscripts)

Take-aways

- Be **transparent** about your full research workflow: research questions, methods, data, step-by-step procedures and analyses
- Make sure you have good **documentation** for all outputs and all stages of your research process
- Keep track of **versions** and do a solid **quality check** of your methods, data and analyses
- **Verify** your own work: try to reproduce your own results and/or have others do it
- Make your methods, data and analyses **open** (if you can)

ReproducibiliTea

Journal Club

**JOIN IN AND DISCUSS WITH FELLOW
STUDENTS AND RESEARCHERS**

**OPEN RESEARCH, REPRODUCIBILITY
and RESEARCH IMPROVEMENT**



Join us

Everyone is welcome to join us - whether you are an enthusiast of open and reproducible research, a skeptic, or a cautious explorer. Currently, all meetings are hybrid with the possibility of joining on-site at Blindern or via Zoom. Grab a cup of tea (coffee?) and join us!

Subscribe to our mailing list



RNO



NOR



Norway

Reproducibility Network

Welcome to Norway's Reproducibility Network!

Norwegian Reproducibility Network (NORRN) is a peer-led consortium located in Norway. It follows an organisational format adopted internationally as nationwide "Reproducibility Networks". NORRN collaborates with the other Reproducibility Networks whilst remaining a unique and independent community.

[Norwegian version of this page](#)

Digital Scholarship Centre

At the Digital Scholarship Centre (DSC) you get guidance on how you can make the best possible use of digital tools and methods in your research and communication activities.

Open Access →

Information about open access publishing, publisher agreements, self-archiving, requirements, and guidelines.

Open and reproducible research →

Make your research more transparent and reproducible.

Research Data Management →

Managing your data both during and after a research project.

Text-mining →

Information about digital tools for searching, mining, and analysing textual data.

Systematic search →

Information about systematic literature searching, how to get started, and how to get help.

Visualisation →

Use of visual methods to explore, communicate and understand data.

Carpentry@UiO →

Offers workshops in foundational digital skills such as coding and data management.

Reference management →

Styles, tools, and information on reference management.

Open and reproducible research

[Norwegian
version of this
page](#)

Learn about how to make your research more open and reproducible and get involved in initiatives and communities that are interested in sharing and improving research at UiO.

Open research

Research methods

workshop-bilder

More and more researchers and students across disciplines are implementing open research practices, preregistering their hypotheses, methods, and analysis plans and sharing research materials, data and analysis scripts. Digital Scholarship Center can help you learn about and implement these practices in your own research as well as advise on the policies and requirements from funders.

Open Science Lunch →

Every last Thursday of the month we meet at noon to discuss topics related to open research.

ReproducibiliTea@UiO →

Join us for a Journal Club where we read and discuss papers on open research and meta-science.

Norwegian Reproducibility Network →

Join a broader community that aims to promote and enable rigorous, robust and transparent research practices in Norway

Courses and workshops →

Click here for the list of upcoming and previous courses and workshops on open and reproducible research at UiO.

Det senteret for digitalforskerstøttes nyhetsbrev,
en del av Universitetsbiblioteket i Oslo

The Digital Scholarship Centre's Newsletter,
part of the University of Oslo Library

DSC NEWS

Senter for digitalforskerstøtte
Digital Scholarship Centre



<https://sympa.uio.no/ub.uio.no/subscribe/dsc-news/subscribe>

More resources:

The Turing Way: Guide for Reproducible Research

<https://the-turing-way.netlify.app/reproducible-research/reproducible-research.html>

CodeRefinery: Reproducible Research

<https://coderefinery.github.io/reproducible-research/motivation/>



Thank you!



- Be **transparent** about the full research workflow: questions, methods, data, step-by-step procedures and analyses
- Make sure you have good **documentation** for all outputs and all stages of your research process
- Keep track of **versions** and do a solid **quality check** of your methods, data and analyses
- **Verify** your own work: try to reproduce your own results and/or have others do it
- Make your methods, data and/or analyses **open** (if you can)

Agata Bochynska, PhD

Open Research and Digital Scholarship Center
University of Oslo Library

@AgataBochynska

agata.bochynska@ub.uio.no