Code execution and peer review make reproducibility possible

Introduction @ CODECHECK and TU Delft Hackathon

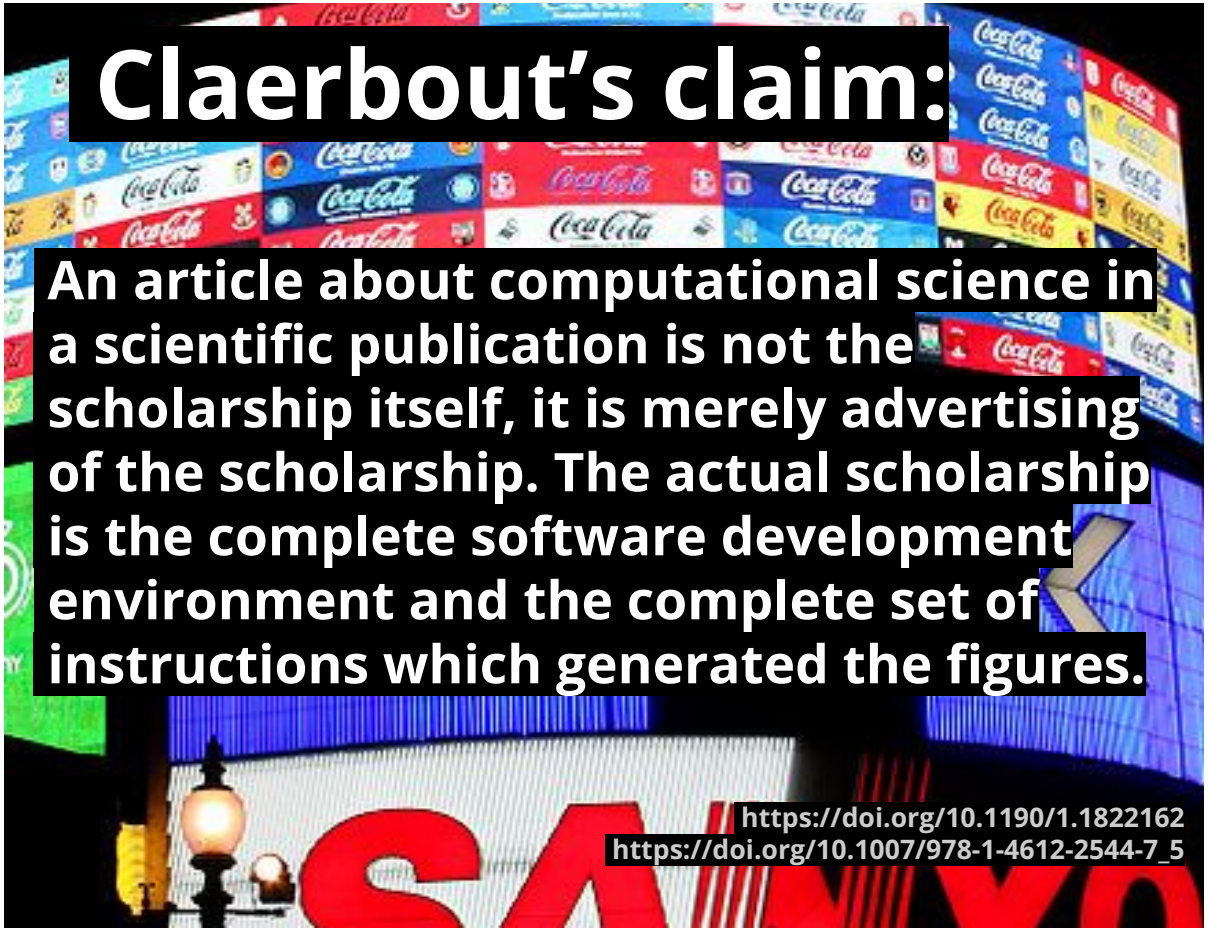
Sep 2023

Daniel Nüst, Stephen Eglen & all CODECHECK supporteers

https://bit.ly/check-delft

Closed and irreproducible research



How to draw an owl

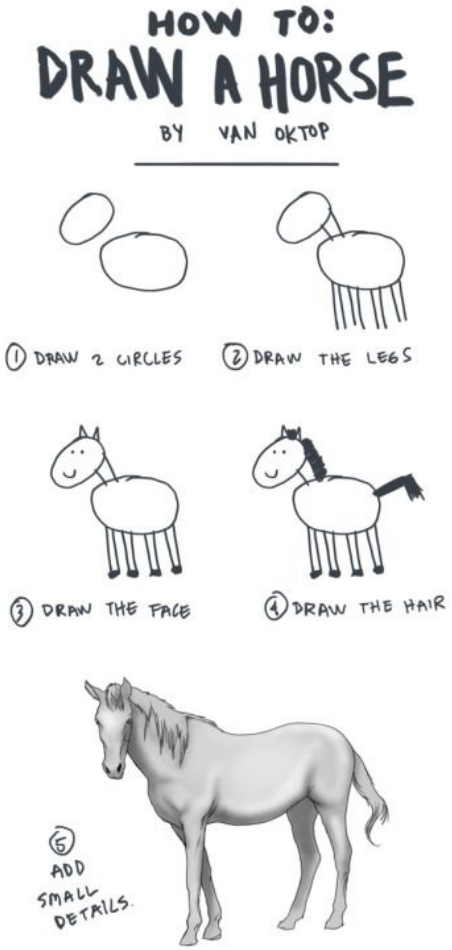
1.

1. Draw some circles

2.

2. Draw the rest of the fucking owl

https://www.reddit.com/r/pics/comments/d3zhx/how_to_draw_an_owl/



https://oktop.tumblr.com/post/15352780846

One thing

Have a README: all else is details.

Show willingness to help, but don't stop publishing because lacking docs. Hard to document for someone else > document for future you, add more on demand.

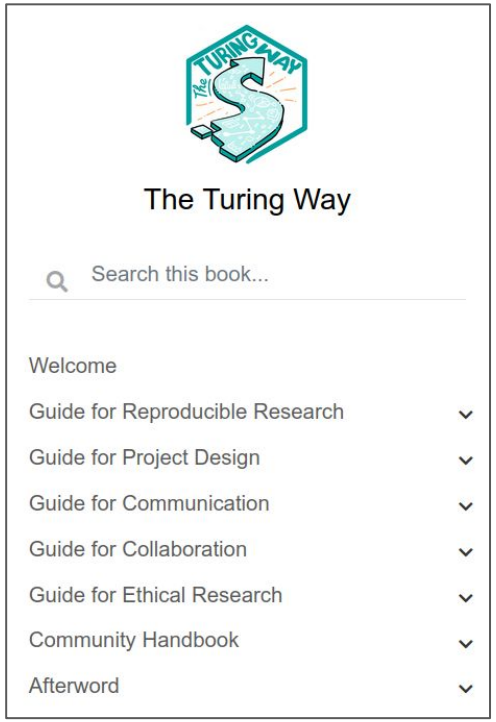
Rule 1 inspired by Greg Wilson's Teching Tech Together (http://teachtogether.tech/en/index.html) Rule 1.

Four things on reproducible research

Have a README: all else is details.
Have a colleague run your workflow before submission.
Reproduce papers (or return the favour 👆).
Publish code and data, cite it.

Rule 1 inspired by Greg Wilson's Teching Tech Together (http://teachtogether.tech/en/index.html) Rule 1.

The Turing Way > https://the-turing-way.netlify.app/

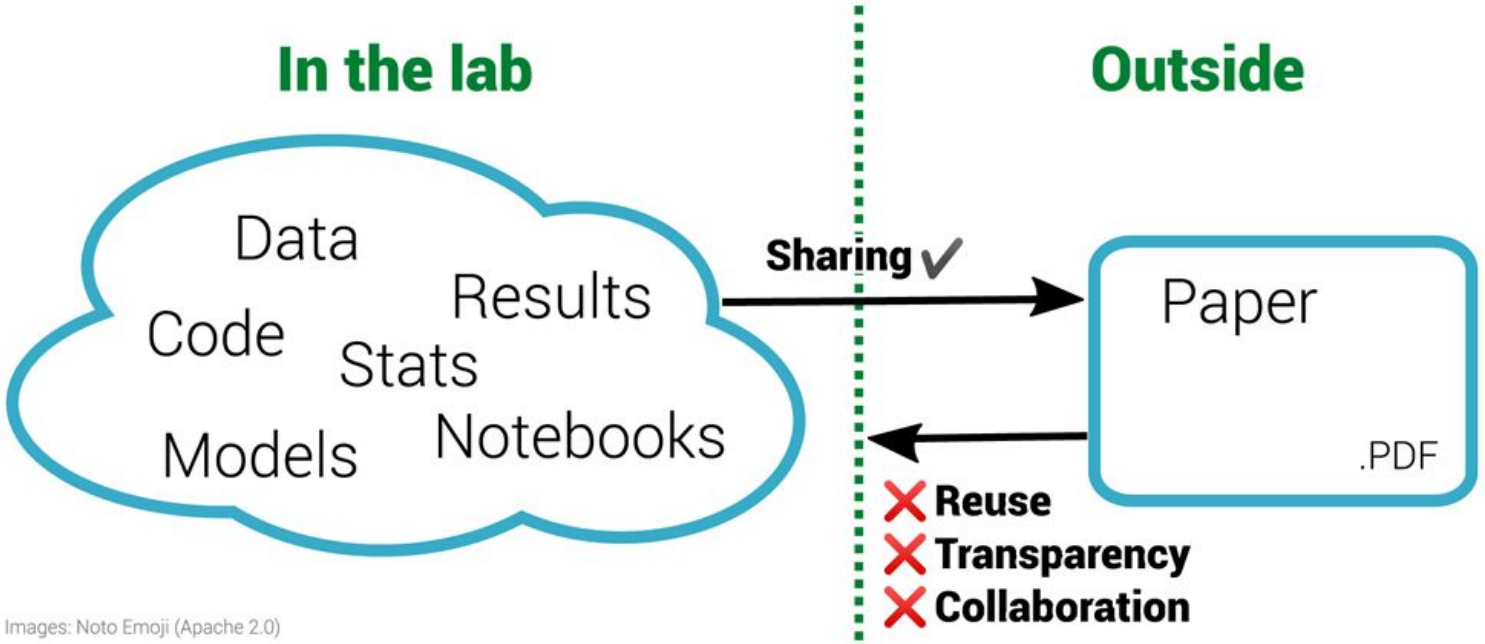


The Turing Way project illustration by Scriberia. Zenodo. http://doi.org/10.5281/zenodo.3332807¶

Reproducible Research Peer Review

Reproducible research and peer review are cornerstones of science. But are they getting along?

https://giphy.com/gifs/suspicious-YNEpBZC0Ly08M https://giphy.com/gifs/lol-futurama-humor-cFgb5p5e1My3K



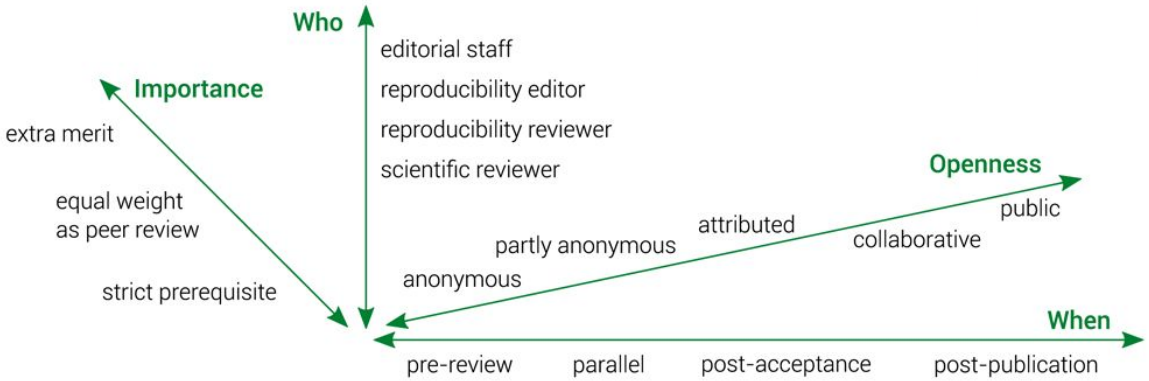
The inverse problem in reproducible research. Figure 1 of https://doi.org/10.12688/f1000research.51738.1

The left half of the diagram shows a diverse range of materials used within a laboratory. These materials are often then condensed for sharing with the outside world via the research paper, a static PDF document. Working backwards from the PDF to the underlying materials is impossible. This prohibits reuse and is not only non-transparent for a specific paper but is also ineffective for science as a whole. By sharing the materials on the left, others outside the lab can enhance this work.

One re-execution of computational workflow by codechecker during peer review

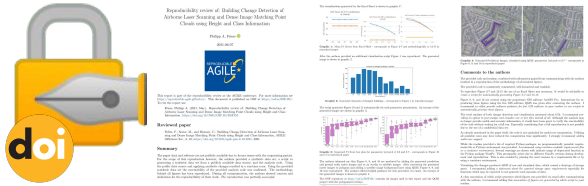
CODE CHECK ✓
https://codecheck.org.uk/

Independent execution of computations underlying research articles.



1. Codecheckers record but don't investigate or fix.
2. Communication between humans is key.
3. Credit is given to codecheckers.
4. Workflows must be auditable.
5. Open by default and transitional by disposition.

CODE WORKS ✓



50+ Certificates
https://codecheck.org.uk/register/

The CODECHECK variations. Figure 1 of https://doi.org/10.12688/f1000research.51738.1

CODECHECK certificates (= AGILE Reproducibility Reports)

Published with a DOI

Title page, cites the paper

Paper links to report via URL/badge (no citation)

Automatically added to ORCID profile

CODECHECK certificate 2020-001

http://doi.org/10.5281/zenodo.3674056

Item	Value
Title	ShinyLearner: A containerized benchmarking tool for machine-learning classification of tabular data.
Authors	Terry J Lee; Erica Suh; Kimball Hill; Stephen R Piccolo
Reference	Paper to appear in Gigascience.
Codechecker	Stephen J. Eglen https://orcid.org/0000-0001-8607-8025
Date of check:	2019-02-14 10:00:00
Summary:	Only visualiation steps performed, rather than machine learning (which could take several hours/days). The created figures match those in the article. The content of other output files was not checked.
Repository:	https://github.com/codecheckers/Piccolo-2020

Table 1: CODECHECK summary

File	Comment	Size
Figures/Datasets_Basic_AUROC.pdf	Figure 2 of manuscript	8078
Figures/Predictions_Histograms.pdf	Figure 3 of manuscript	8727
Figures/Algorithms_ParamsImprovement_AUROC.pdf	Figure 4 of manuscript	7837
Figures/Algorithms_FSImprovement_AUROC.pdf	Figure 5 of manuscript	8190
Figures/FS_vs_CL.pdf	Figure 6 of manuscript	6521
Figures/FS_NumFeatures.pdf	Figure 7 of manuscript	5810
Tables/Basic_DiffFromMedian.tsv	Example output table 1 (not in manuscript)	619
Tables/ParamOpt_Improvement.tsv	Example output table 2 (not in manuscript)	14216

Table 2: Summary of output files generated

Summary

The reproduction of the figures in the manuscript was straightforward given that the authors provided a Rmarkdown document that processed the results data files. The results data files were not independently reproduced at this stage because of the long compute time.

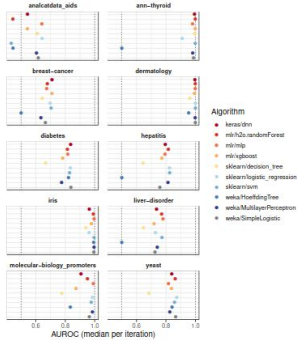


Figure 1: Figure 2 of manuscript

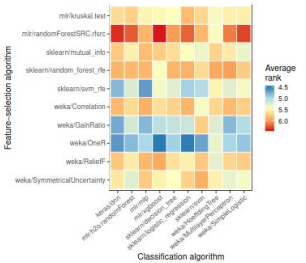


Figure 5: Figure 6 of manuscript

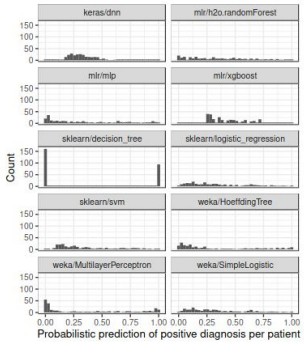


Figure 2: Figure 3 of manuscript

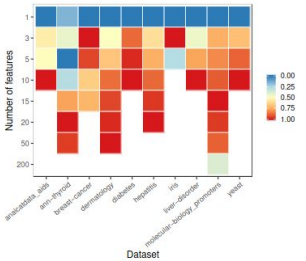


Figure 6: Figure 7 of manuscript

http://doi.org/10.5281/zenodo.3674056

CODECHECK Certificate 2022-018

10.5281/zenodo.7084333

Ranicre Silva

September 27, 2022

Table 1: CODECHECK summary

Title	svaRetro and svaNUMT: Modular packages for annotation of retrotransposed transcripts and nuclear integration of mitochondrial DNA in genome sequencing data
Authors	Ruining Dong, Daniel Cameron, Justin Bedo, Anthony T Papenfuss
Reference	https://doi.org/10.46471/gigabyte.70
Summary	Only visualisation steps performed. All created figures match those in the article.
Repository	https://gitlab.com/cdchck/community-codechecks/2022-svaRetro-svaNUMT.git

Table 2: Summary of output files generated

Files	Comment
figure-2b.pdf	Figure 2(b) of the article
figure-3b.pdf	Figure 3(b) of the article
figure-4.pdf	Figure 4 of the article
figure-5.pdf	Figure 5 of the article
figure-6.pdf	Figure 6 of the article

1

Summary

The reproduction of the figures, from ou
R Markdown (`.Rmd`) files. Figure 3 ar
reproduced!

CODECHECKER notes

Data and Code

As a repository was not provided by
`MANIFEST` was created. `Scripts.zip`
supplemented material in Zenodo (Dor

```
$ make download
```

Software Installation

The provided `.Rmd` files requires many
`bioconductor/bioconductor_docker`
by running

```
$ docker compose up dev
```

Packages installation instructions are i

Running the Script

`Figures2-4.Rmd` is the main script an
history for details. To regenerate the fi
as part of the Bioconductor Docker im

Figure 4 uses statistics from `sim_read`

`gnomad.Rmd` is the script that renders F
at the end of the document, see Git his

```
#function from SVEnsemble
wkdir <- getwd()
gnomad.bnd.gr <- suppressWarning

gnomad.rt <- rtDetect(
  filter(gnomad.bnd.gr, FILTER=="
  hg19.genes,
  maxgap = 1000,
  minscore = 0.4
)
```

takes a couple of hours to execute and

```
gnomad.insSite.pass.rmsk <- gnom
  find_overlaps(., rmsk.gr, maxg
  unique() %>%
  as_tibble() %>%
  count(repClass) %>%
  bind_rows(tibble(repClass='non
```

failed with

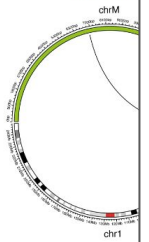
```
Error in count(., repClass) : Ar
```

that was resolved by replacing

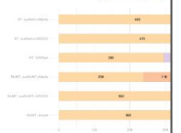
```
count(repClass) %>%
```

Fi

(a) From pre-pr



(a) From pre-pr



References

Dong, Ruining, Daniel Cameron, Justin Bedo, and Anthony T Papenfuss. 2022. "Data and Scripts for the Manuscript of svaRetro and svaNUMT: Modular Packages for Annotating Retrotransposed Transcripts and Nuclear Integration of Mitochondrial DNA in Genome Sequencing Data." Zenodo. https://doi.org/10.5281/ZENODO.7006177.

Colophon

This document was built with Quarto.

Session Info

```
sessionInfo()
```

```
R version 4.2.1 (2022-06-23)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 22.04.1 LTS

Matrix products: default
BLAS:   /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.10.0
LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.10.0

locale:
 [1] LC_CTYPE=en_GB.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_GB.UTF-8        LC_COLLATE=en_GB.UTF-8
 [5] LC_MONETARY=en_GB.UTF-8    LC_MESSAGES=en_GB.UTF-8
 [7] LC_PAPER=en_GB.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_GB.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

loaded via a namespace (and not attached):
 [1] digest_0.6.29   jsonlite_1.8.0  magrittr_2.0.3  evaluate_0.16
 [5] highr_0.9       rlang_1.0.5     stringi_1.7.8   cli_3.4.0
 [9] rstudioapi_0.14 rmarkdown_2.16  tools_4.2.1     stringr_1.4.1
[13] xfun_0.33       yaml_2.3.5      fastmap_1.1.0   compiler_4.2.1
[17] htmltools_0.5.3 knitr_1.40
```

Figures2-4.Rmd's session info:

```
## R version 4.2.1 (2022-06-23)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 20.04.4 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3
## LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/liblapack.so.3
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
```

6

https://doi.org/10.5281/zenodo.7084333

What can you do today?

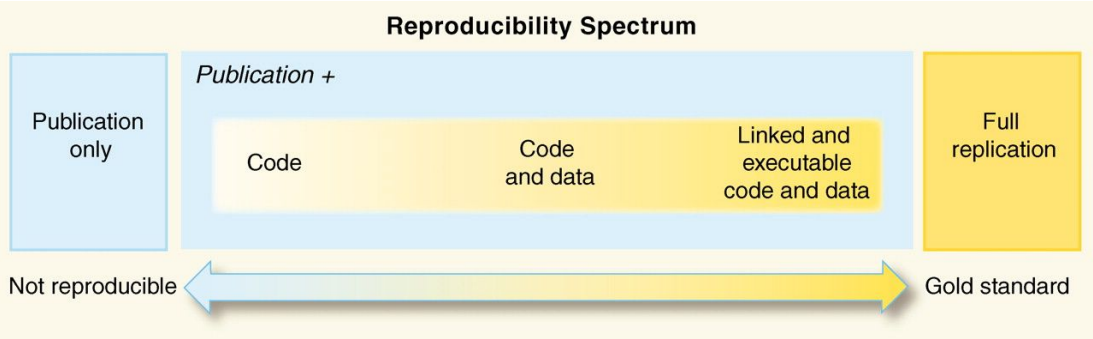
Learn how to

CODECHECK

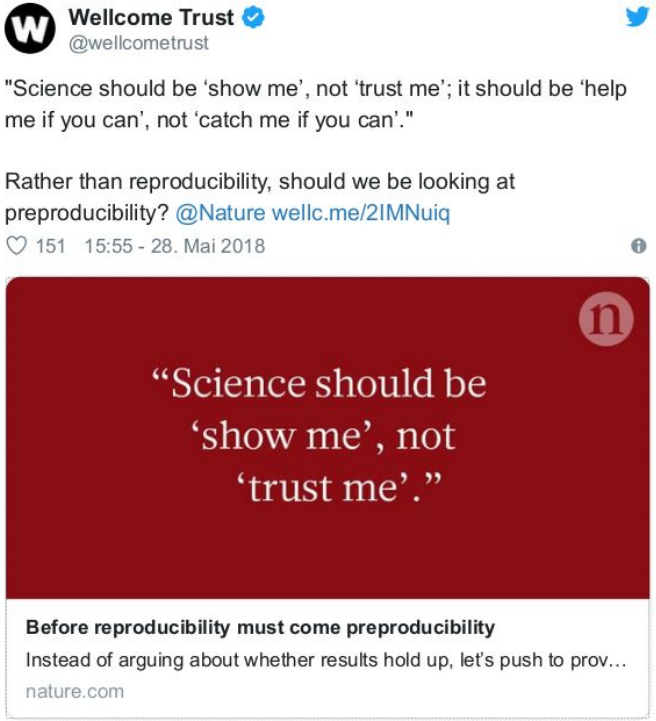
and join a local community of codecheckers
to shift practice one paper at a time.

https://codecheck.org.uk/

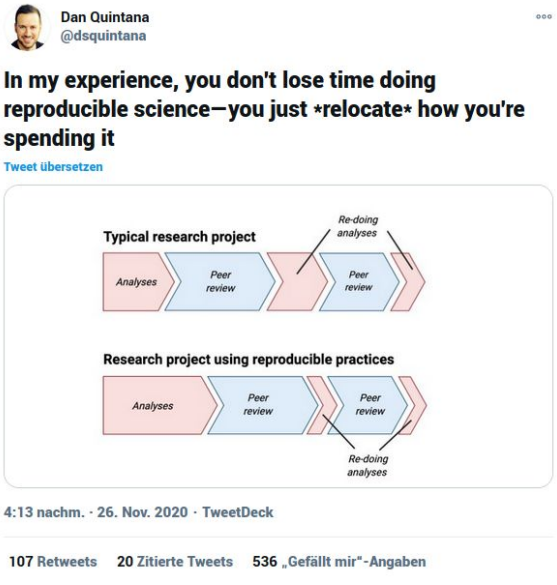
Reproducible Research & Open Science

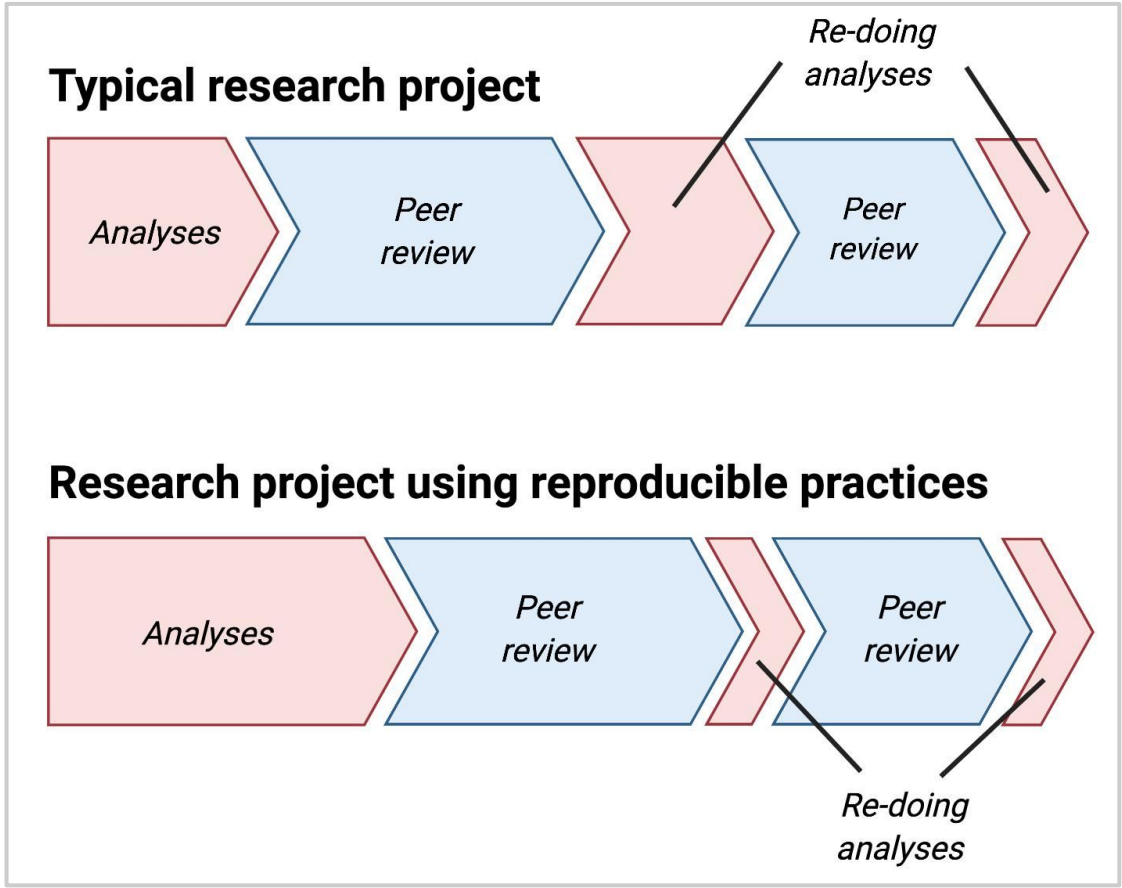


https://doi.org/10.1126/science.1213847



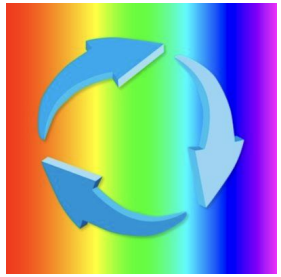
https://www.nature.com/articles/d41586-018-05256-0

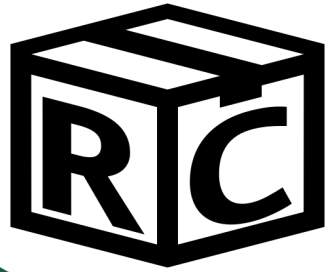




Quintana, D. S. (2020, November 28). Five things about open and reproducible science that every early career researcher should know. https://doi.org/10.17605/OSF.IO/DZTVQ

What can scientists do?





Take one step at a time.

Create and publish Research Compendia (Your code is good enough!): https://research-compendium.science/

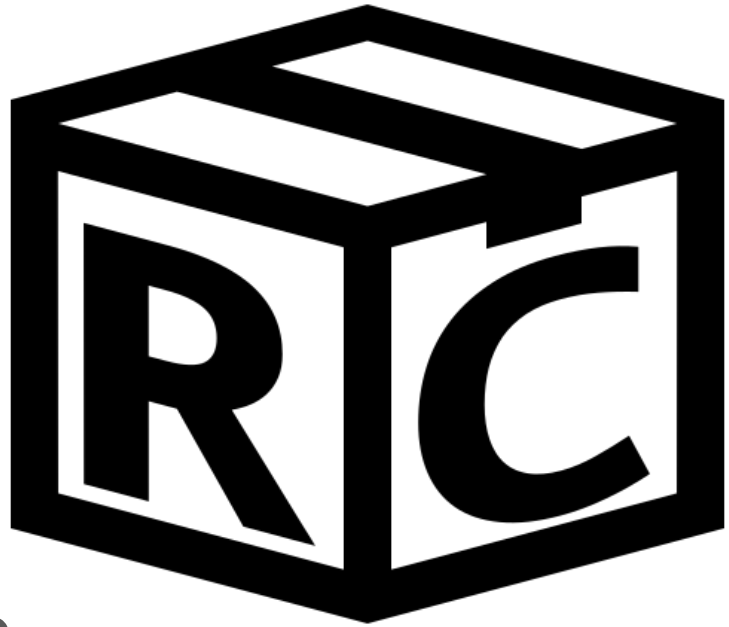
Become a codechecker or reprohacker.
Join a Reproducibility 4 Everyone workshop.

Strive to be an open science champion especially if you're junior in your field. [RIOT talk by Gavin Buckinham; preprint by Sam Westwood]

Be the change, find communities, do not rely on those in power - they don't know!

Research Compendia

= programming language packaging + science stuff
= templates
= community practices (lab, discipline, language, method)



research-compendium.science

What can communities and institutions do?

Introduce reproducibility reviews - CODECHECK (or not) - at your journals, labs, collaborations!

Workshops on RCR, ReproHacks

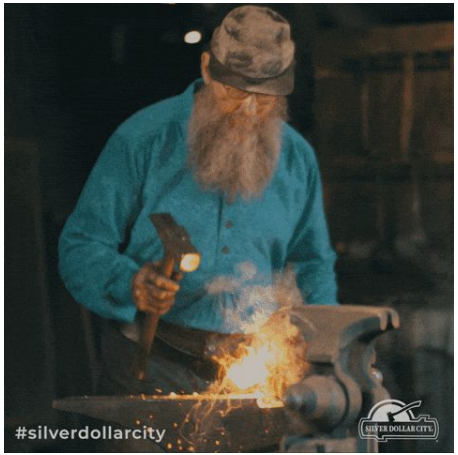
Provide support

Establish rewards and incentives

Enable community discourse

Awareness > Change

https://giphy.com/gifs/chicagodancecrash-KCqjrcPfL55q3MkgHZ



Learn crafts by doing. Be kind. Help.