

How (and why) to work as an open scientist in the 21st century

Stephen J Eglén
<https://sje30.github.io>
sje30@cam.ac.uk

Cambridge Computational Biology Institute
University of Cambridge
@StephenEglén

Slides: <http://bit.ly/eglen-todos> (CC-BY license)

Declarations

1. Affiliate editor of *bioRxiv*
2. Senior editor of *Scientific Data*

Acknowledgements

Laurent Gatto, Corina Logan, Daniel Nüst, Ben Marwick

About me

My research and how I got into open science.

What is open science

- My definition of open science: a way to approach research which promotes sharing of (all / most) research artifacts.
- From this, open access (OA) to research papers is normally assumed. But we all make mistakes...

Ooh, a Nature publication?

Sharing your data is easier than you think

Stephen Eglen 

Nature 510, 340(2014) | [Cite this article](#)

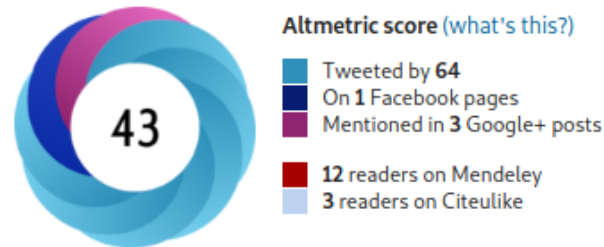
457 Accesses | 3 Citations | 43 Altmetric | [Metrics](#)

Geoffrey Goodhill questions some of the practicalities of open data-sharing policies (*Nature* 509, 33; 2014), but I believe that his concerns are largely unfounded.

Storing large volumes of raw data is costly, but many items destined for sharing are highly processed and relatively small. The mouse-brain connectome, for example, is available as a 3-megabyte file derived from many gigabytes of raw data (S. W. Oh *et al.* *Nature* 508, 207–214; 2014). Neither is there a shortage of repositories: many institutional databases are freely available and well supported (such as zenodo.org, maintained by CERN, Europe's particle-physics lab in Geneva, Switzerland). More repositories will come online as researchers learn how to share data more effectively.

Contrary to Goodhill's suggestion, sharing computer code does not necessarily demand much time investment (see, for example, D. C. Ince *et al.* *Nature* 482, 485–488; 2012). Code is a valuable part of a paper, so everyone benefits if its authors assume from the start that it will be shared or reused. Also, people releasing code are under no obligation to maintain it.

Online attention




This Altmetric score means that the article is:

- in the 96th percentile (ranked 6,798th) of the 189,057 tracked articles of a similar age in all journals
- in the 53rd percentile (ranked 408th) of the 881 tracked articles of a similar age in *Nature*


Twitter can be a tough place ...

Altmetric Details Page

 **Marc Perry**
@mdperry


Lol, what an irony! <http://t.co/yKZdo2Qq2Y> <http://t.co/iAyQENqVpM>

22 Jun 2014

 **Daniel MacArthur**
@dgmacarthur


Lol, what an irony! <http://t.co/yKZdo2Qq2Y> <http://t.co/iAyQENqVpM>

22 Jun 2014

 **Petr Danecek**
@petrdanecek

Lol, what an irony! <http://t.co/yKZdo2Qq2Y> <http://t.co/iAyQENqVpM>

22 Jun 2014

 **INCF**
@INCForg

INCF UK Node's Stephen Eglén in Nature: Sharing your data is easier than you think <http://t.co/XKP9g2UQMX>

22 Jun 2014

from which I learnt altmetrics can mislead and that I should **preprint**.

The reproducibility crisis

Many key findings in publications are either not independently verified, or fail verification when it is attempted (Baker, 2016).

Duke oncogenomics scandal. Awesome detective work by Keith Baggerley and Kevin Coombes. <https://www.youtube.com/watch?v=7gYIs7uYbMo>

Disclaimer: do I mean "reproducibility" or "replicability"? (Barba 2018)
<https://arxiv.org/pdf/1802.03311.pdf>

Reproducibility crisis

Crisis? What crisis? Science is getting better, right?

"Negative results are disappearing from most disciplines and countries" ([Fanelli, 2012](#)).
Study of ~4600 papers:

Negative results are disappearing from most disciplines and countries

897

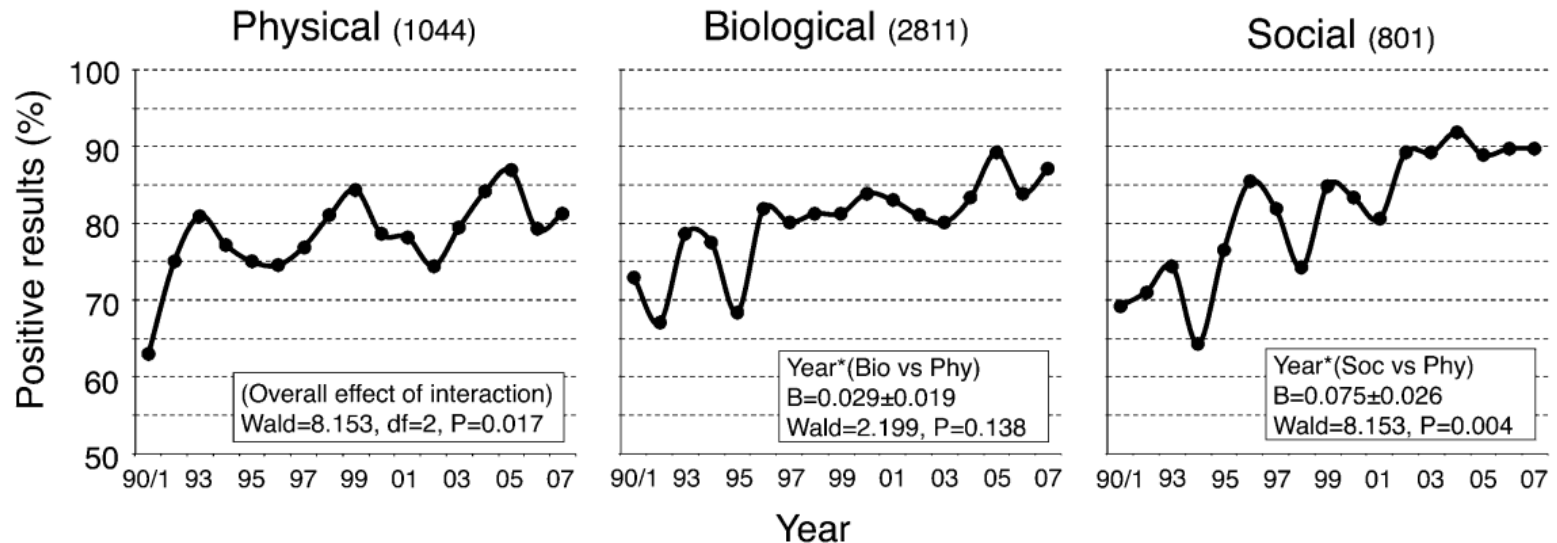
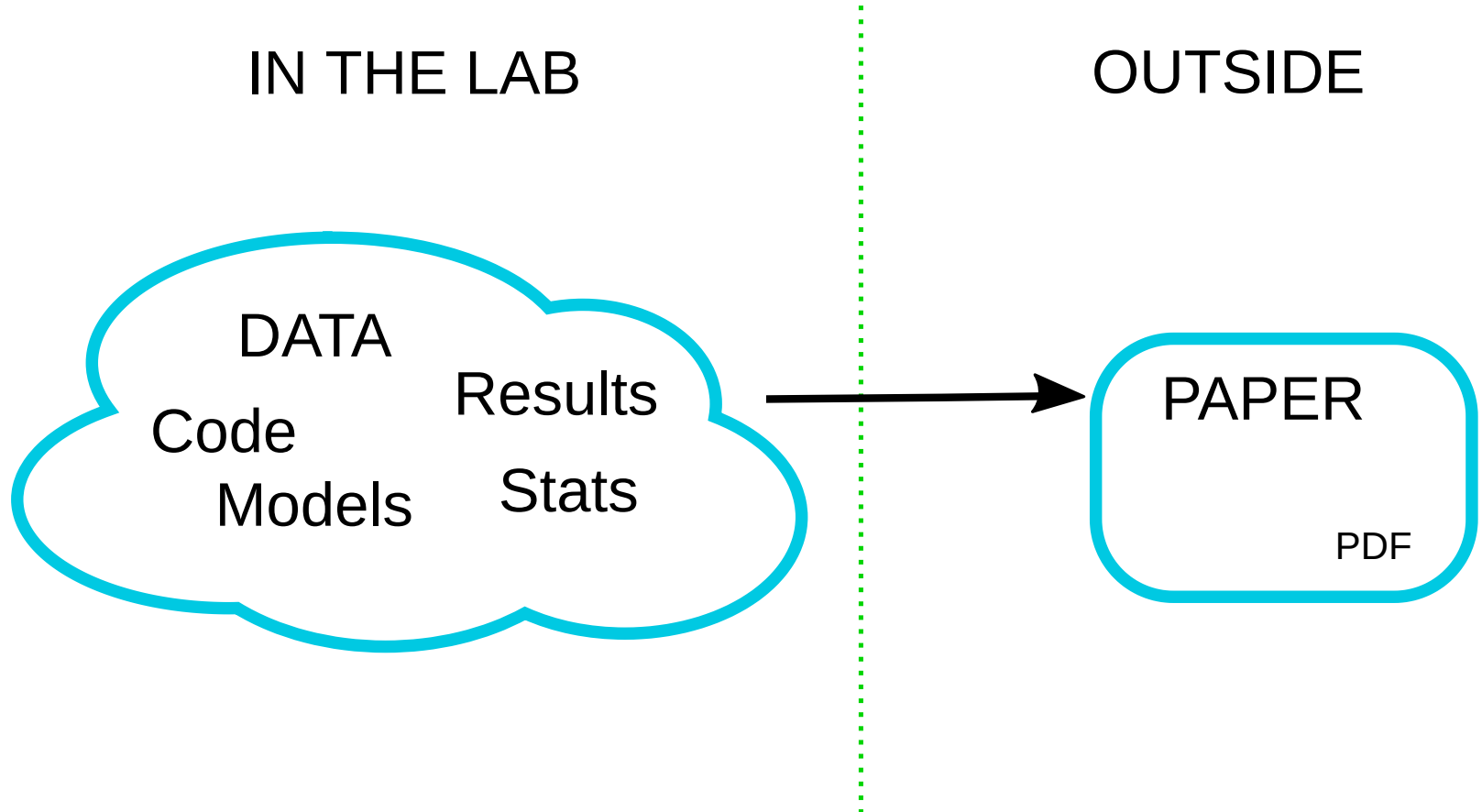


Fig. 3 Percentage of papers reporting a support for the tested hypothesis plotted against year of publication and divided by scientific domain of the journal (physical, biological and social sciences). Logistic regression estimates are interaction effects in a hierarchically well-formulated model. The main effects of this model, calculated with interaction components removed, are reported in Table 1. Numbers in brackets are sample size

WHY

Science as an inverse problem



Why share your work?

An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.

[Buckheit & Donoho \(1995\)](#)

The problem is that most modern science is so complicated, and most journal articles so brief, it's impossible for the article to include details of many important methods and decisions made by the researcher

[Marwick \(2015\)](#)

This often plays into the moral argument for sharing your work (taxpayers money etc), but doesn't apply (e.g. Humanities).

Moral or selfish approach to sharing?

Markowitz *Genome Biology* (2015) 16:274
DOI 10.1186/s13059-015-0850-7



COMMENT

Open Access

Five selfish reasons to work reproducibly



Florian Markowitz

Abstract

And so, my fellow scientists: ask not what you can do for reproducibility; ask what reproducibility can do for you! Here, I present five reasons why working reproducibly pays off in the long run and is in the self-interest of every ambitious, career-oriented scientist.

Keywords: Reproducibility, Scientific career

how science actually is. And, whether you like it or not, science is all about more publications, more impact factor, more money and more career. More, more, more... so how does working reproducibly help me achieve more as a scientist.

Reproducibility: what's in it for me?

In this article, I present five reasons why working reproducibly pays off in the long run and is in the self-interest of every ambitious, career-oriented scientist.

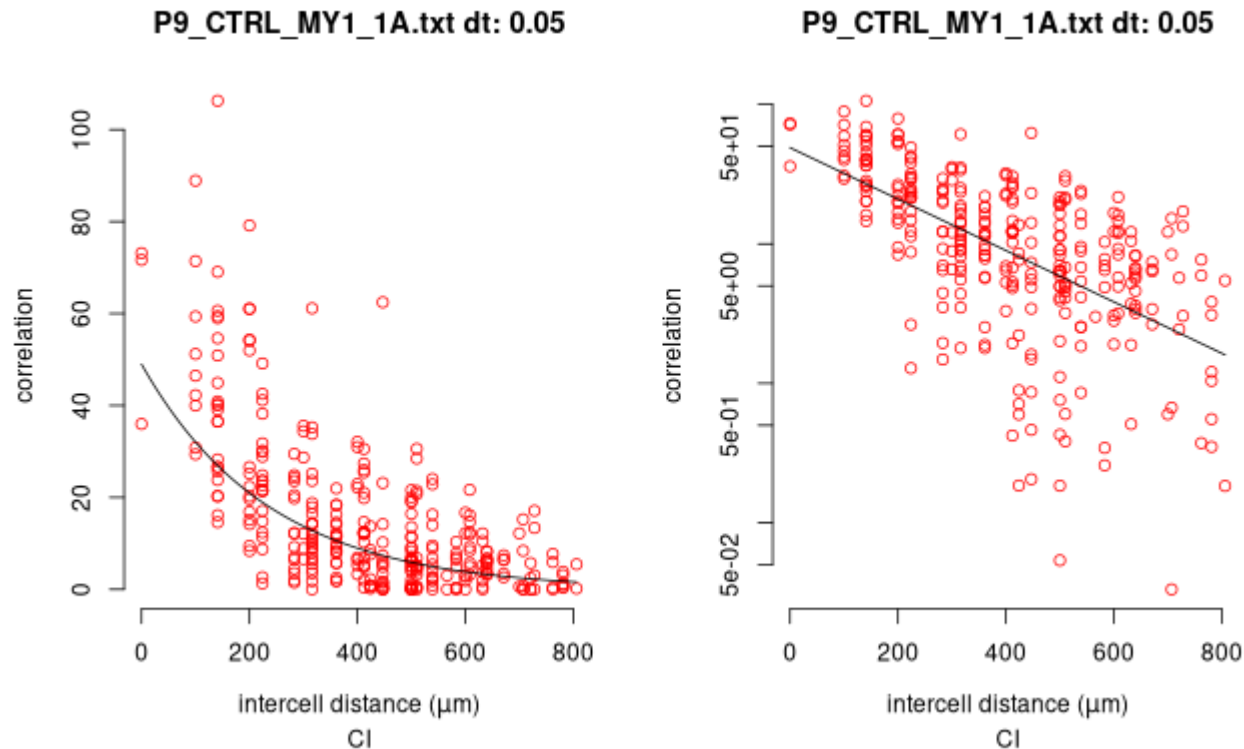
Selfish reasons to share

Why not align what is good for science with what is good for scientists?

1. Funding mandates (REF + enforcement from Wellcome Trust)
2. Credit through data papers
3. Fixes data bugs / errors in analysis
4. Prevent data loss ([Vines et al 2014](#)), e.g. students have a habit of leaving...
5. Your future self is probably one of the main beneficiaries of sharing.
6. Now is a very good time to be an open scientist.
7. Leads to further collaborations
8. Reviewers can do more work...

Reviewers doing your work

I would use an ordinate log scale for this bottom right panel (as done in Fig. 3). But since the authors gave me everything, I can do it! by redefining fourplot as follows:



HOW

Rule 1: Data should be shared

Given the cost of generating data, I think data relating to a publication should be shared along with the paper.

Funders (and increasingly many, but not all, journals) agree.

Exceptions: human / clinical datasets where GDPR relevant.

Where to store your research data?

Ask your local librarian experts about the possibilities available to you:

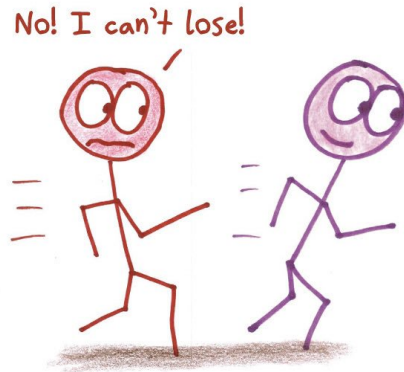
- Institutional repositories (local support but what if you move)
- National repositories (probably more flexible)
- Subject-specific repositories: <https://journals.plos.org/plosone/s/data-availability#loc-recommended-repositories>.
- The big general repos: [Zenodo](#) [figshare](#)

Learn about licensing

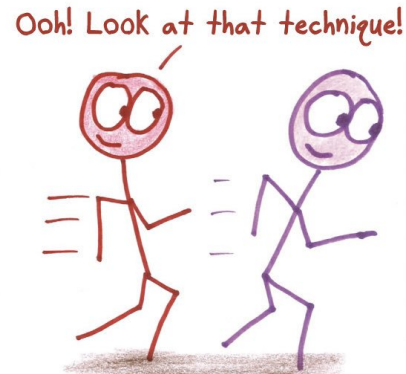
- How can you licence your work so that you can share it and yet protect it?
- Again, best to seek local advice from librarians.
- <https://choosealicense.com/non-software/>
- My preferences: MIT (for code) and CC BY (for data etc).

Get a good mentor

A good mathematician
wants to be the best.



A great mathematician
wants to learn from
the best.



Science is hard enough without having to be competitive.

Having a local group to talk about these issues is great. I have learnt much about open science from postdocs (often online).

Format-free submissions

Julian Budd makes simple point clearly: let's not waste time reformatting our papers on first submission.

[nature.com/nature/journal ...](https://www.nature.com/nature/journal)

NATURE | CORRESPONDENCE



Publishing: Reformatting wastes public funds

Julian Budd

Nature **543**, 40 (02 March 2017) | doi:10.1038/543040e
Published online 01 March 2017

 **PDF**  **Citation**  **Reprints**  **Rights & permissions**  **Article metrics**

Subject terms: [Publishing](#) · [Peer review](#)

Limited public funds for scientific research are being spent on reformatting manuscripts for different journals, without any apparent gain for science or society (see [Q. Guo *Nature* 540, 525; 2016](#) and [J. P. Moore *Nature* 542, 31; 2017](#)). As a peer reviewer, I am interested in a manuscript's content — not its format. The increasing popularity of preprint servers indicates that format does not bother readers either.

In 2013, for instance, *Nature* published less than 8% (856 of 10,952) of the research papers submitted (see go.nature.com/2m102lb). If it takes authors, say, an hour or more to reformat each rejected article for submission to another journal, this will amount to some 10,000 scientist-hours over just one year.

For many papers that are rejected without review, there will be no need to change the scientific content before resubmission, and so paid time spent on reformatting them is not even scientifically justified.

[Source](#)

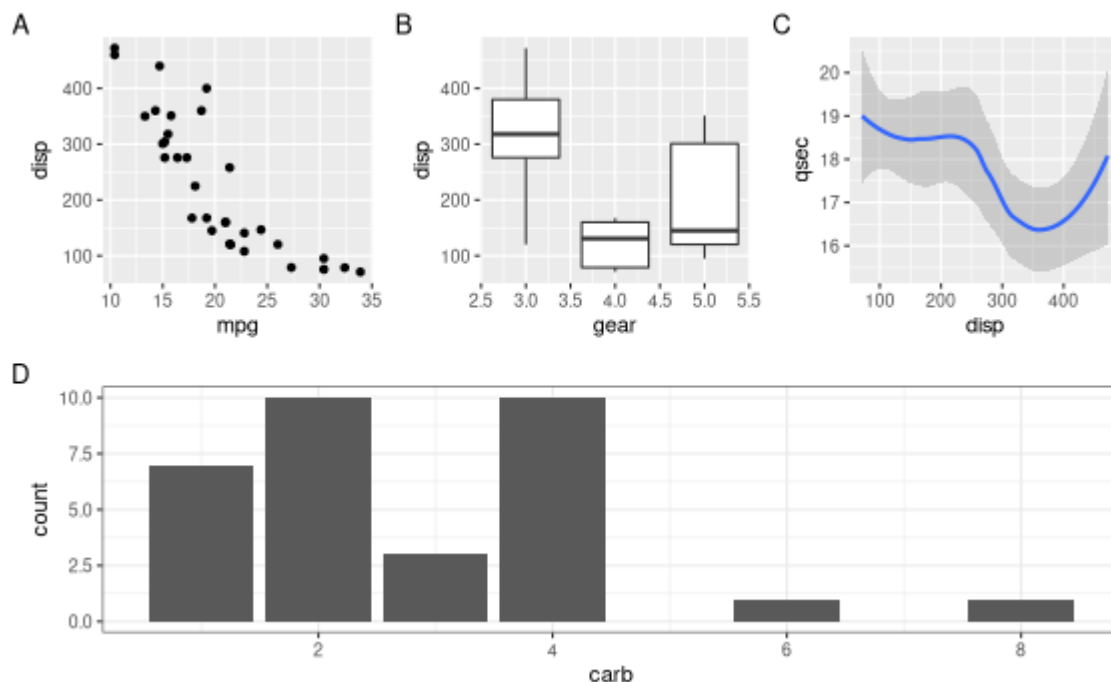
[Don't format manuscripts: Brischoux & Legagneux](#)

Coding

- Coding is a 21st century skill that everyone needs to know.
- Start by moving away from Excel to R for statistical analysis and drawing figures.
- Good Places to learn coding and related skills: [Software Carpentry](#). [Data carpentry](#). [Library Carpentry](#).

Bonus - Reproducible figures

```
library(ggplot2); library(patchwork) # github.com/thomasp85/patchwork
p1 = ggplot(mtcars) + geom_point(aes(mpg, disp)) + labs(tag="A")
p2 = ggplot(mtcars) +
  geom_boxplot(aes(gear, disp, group = gear)) + labs(tag="B")
p3 = ggplot(mtcars) + geom_smooth(aes(dis, qsec)) + labs(tag="C")
p4 = ggplot(mtcars) + geom_bar(aes(carb)) + labs(tag="D")
((p1 | p2 | p3) / p4) + theme_bw()
```



Registered reports

- A positive approach to reducing p-hacking and HARKing (Hypothesising After Results Known). [Cerebral Cortex 2013](#)
- Peer review process split into two steps.
- **Step 1** methods and proposed analyses are pre-registered and reviewed.
- If peer reviews are favourable, authors get "accept in principle" for their paper *regardless* of results.
- **Step 2** after results collected, paper is written.
- This commits authors to think about analysis before experiment
- Extra analyses can be reported as "incidental findings".
- Analysis so far suggests an increase in the reporting of null results, cf earlier slide. <https://www.nature.com/articles/d41586-018-07118-1>
- Reviewers are now much more helpful, commenting on what you should do, rather than what you did wrong.

Preprints



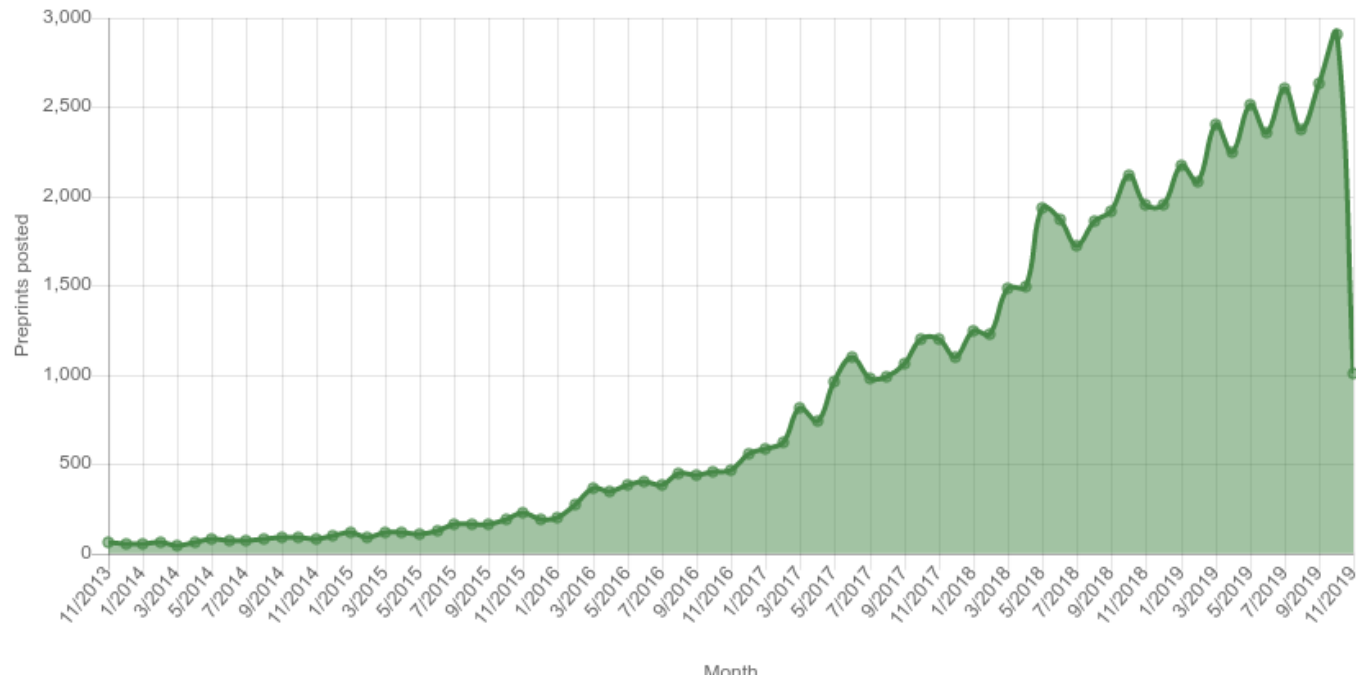
Rxivist combines preprints from [bioRxiv](#) with data from Twitter to help you **find the papers being discussed** in your field. Currently indexing **64,934 bioRxiv papers** from **287,775 authors**.



Site-wide metrics

The numbers below represent the metrics for all papers hosted on [bioRxiv.org](#), based on our indexing of the website.

Monthly submissions, overall



Why preprint?

Advantages (<https://www.plos.org/why-preprint>):

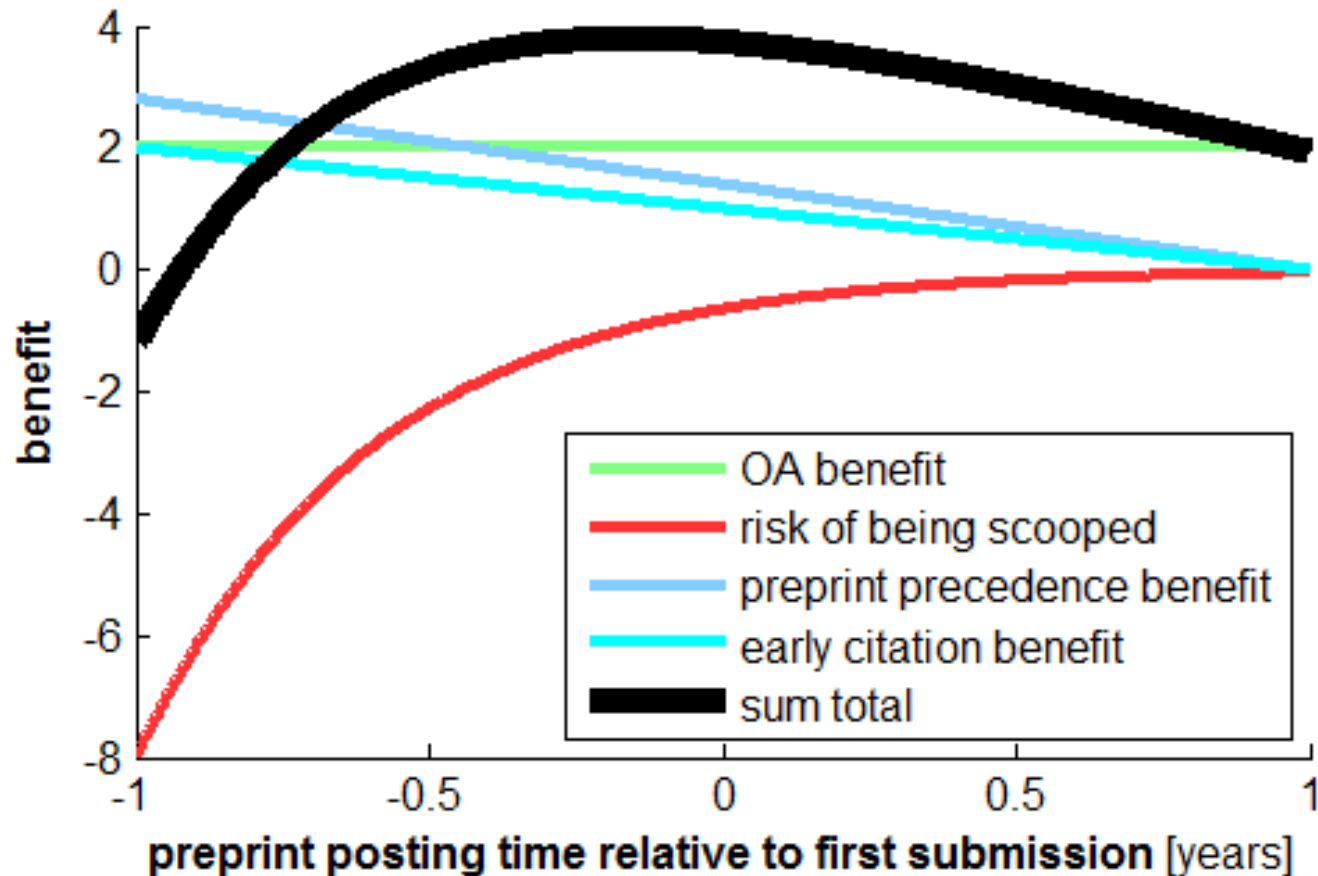
1. Rapid dissemination of results
2. Establish priority
3. Increased attention/citations
4. Career advancement
5. Community
6. Unlimited / timely updates

Also: potential for early feedback, error reduction, improved citations, editors might fish for papers.

Preprints are valid research outputs for REF2021 [Naomi Penfold](#).

When to submit a preprint?

Estimates from <https://nikokriegeskorte.org/2016/03/13/the-selfish-scientists-guide-to-preprint-posting/>



Selfish approach, adopted by many: submit preprint at time of submissions.

Evaluate journal subscription model

Think about which journals you support, by sending papers and reviewing.

Support OA journals as authors and reviewers. (Ever asked a journal for permission to republish one of your own figures?)

Check out status of journal via Sherpa/Romeo service.

What can your institutions & funders do?

- sign up to DORA (Declaration of Research Assessment) (Wellcome Trust have solved this problem). Signing is easy, enforcing is hard.
- Give ECR a voice within institutions.
- Recognise all research artefacts, not just papers, as valuable outputs.
- Give academics more ownership of funds. "Do you really want to spend 5K on an APC, versus other uses in your lab?"
- Routes to OA need to be generate less admin for librarians / funders, not more. Green OA achieves that.
- Support diamond OA initiatives led by academics and academic societies.

Summary

- Find the selfish reasons to make your research reproducible.
- Adopt good practices to help you on your way.
- Writing code and analysing data in groups can be very motivating.
- Use new tech if you want, but old tech works too.

Conclusions

- Technical challenges << Societal challenges
- You are the future and so is open science!
- Its a journey, we learn these techniques over time.

Extra resources

1. [Turing Way handbook; "a lightly opinionated guide to reproducible data science"](#)
2. [Towards standard practices for sharing code in neuroscience](#)
3. [Ten simple rules for taking advantage of git and github](#)