A practical guide to working reproducibly

OHBM Australia, May 2020

Ben Fulcher

So you're all on board with open science

It is unambiguously good for progressing knowledge and building trust in science
(at a time in history when the public has few institutions left to trust)

So now **you want to act** on this?

"How do I incorporate open science practices into my workflow?"

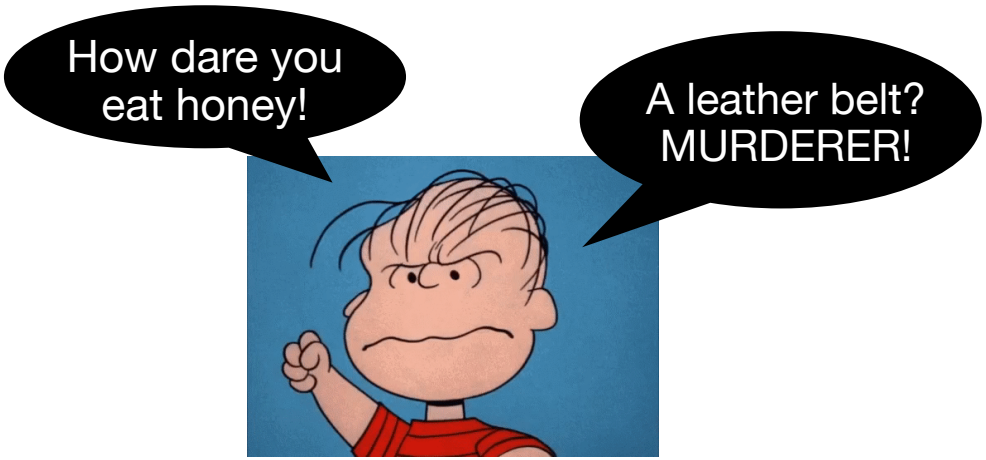
It’s not all or nothing.

Building new habits is hard (and so is doing science)

Any little step you make in the direction of open science is progress

Don’t be intimidated by hard-liners; Make progress little by little, the best you can

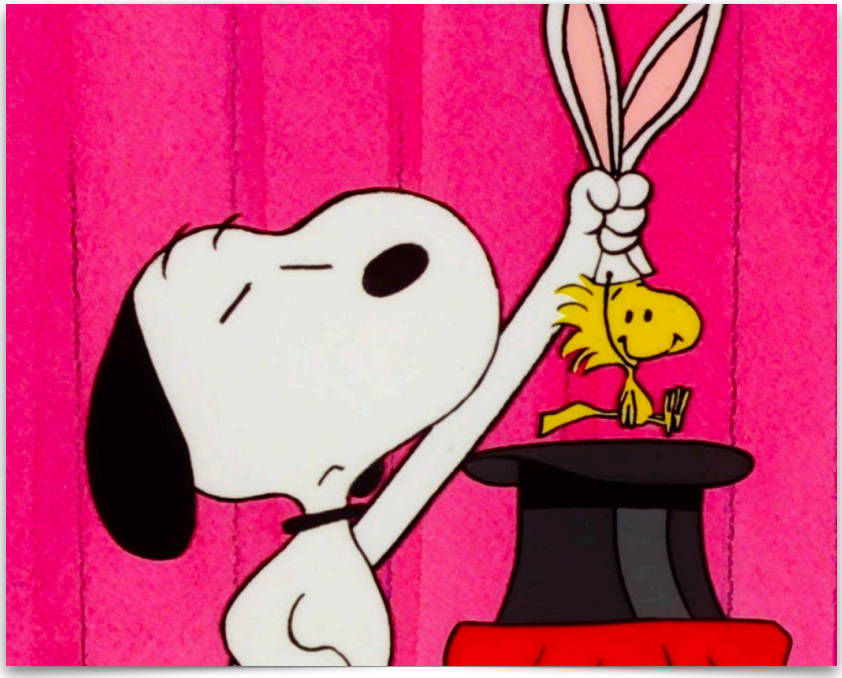
“I want to reduce suffering of animals”



“I want to do open science”

What are we trying to do

“If others cannot **easily reproduce exactly what you’ve done**, then you are not contributing science, you are advertising it.”

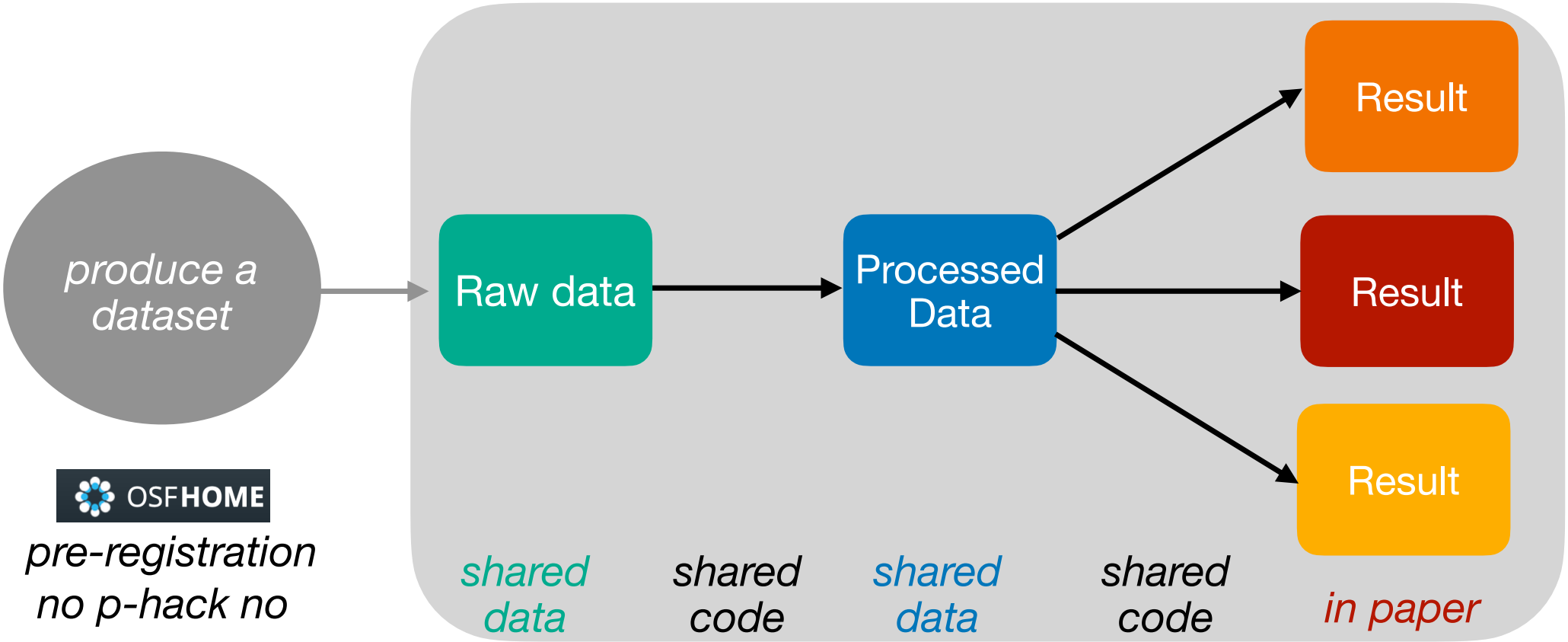


~Donohoe

https://reproducibleresearch.net/

What are we trying to do

"If others cannot **easily reproduce exactly what you've done**, then you are not contributing science, you are advertising it."

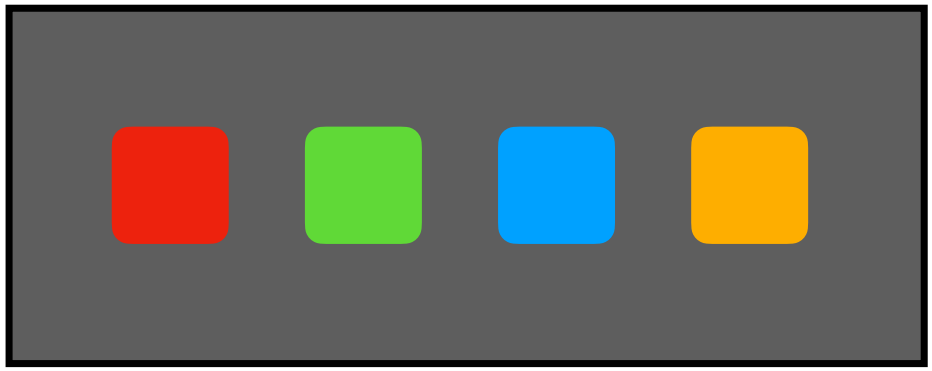


Level 0

You fully describe exactly what you've done

Your description is so clear that any scientist in your field could unambiguously reproduce your results

The only option for GUIs but it is **very hard** to, in words, fully and unambiguously describe even simple analyses



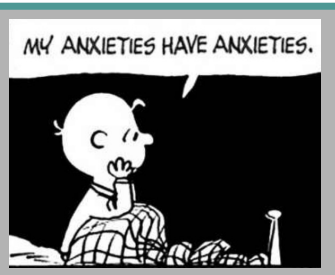
"I loaded data, then pushed the green button and then the red button twice"

Level 1

You provide code that reproduces all of your results.

Complex, multistep analyses are described unambiguously and can be reproduced exactly.

```
dataset1.load
greenButton.push()
redButton.push()
redButton.push()
figure.export
```



Any code is better than none

Levelling Up Your Code's Availability

"Code available on GitHub"

"See supplementary file: `MyCode.zip`"

"Code and data available from the authors on request"

Levelling Up Your Code's Reproducibility

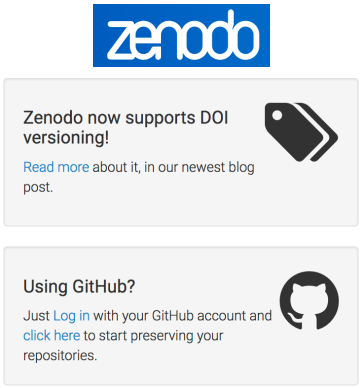
"Code available on GitHub"

But what if code changes after publication?

Code should evolve, but it's important to snapshot the version that was used for a paper

Tag a release, *e.g., v1.0* and mention it in paper (we trust you not to modify)

Give your code a DOI on (e.g.) Zenodo



Levelling Up Your Code's Reproducibility

"Code available on GitHub"

What if software changes?

"It worked a few years ago on some combination of older software"

Fully describe all relevant software, packages, and dependencies (+ Operating System/hardware?)

Set up the full environment

https://www.docker.com/

Levelling Up Your Code's Reproducibility

"Code available on GitHub"

Paid software with proprietary algorithms are a barrier to reproducibility

$$$

Pay to run code

Proprietary algorithms cannot be tested

Free to run code

Open algorithms

"You should join a richer Uni"

"Everything good in life is free"

Levelling Up Your Code's Clarity

No use for reproducibility if code is shared but not clearly documented

Coding for research can be a messy beast as you chase down leads, go down dead ends, etc. (compounded by insufficient training in coding)

Give clear instructions of how to reproduce (e.g., in the `README.md`)

A go-to rule:

Does it pass the person-next-to-me-at-a-conference test?

benfulcher / **mouseGradients**

Unwatch 3 | Star 5 | Fork 3

Code | Issues 0 | Pull requests 0 | Actions | Projects 0 | Wiki | Security 0 | Insights | Settings

Code to reproduce analyses of mouse cortical gradients

Edit

geneexpression | neuroscience | mouse | Manage topics

4 commits | 1 branch | 0 packages | 0 releases | 1 contributor

Branch: master | New pull request | Create new file | Upload files | Find file | Clone or download

benfulcher Trying to get a couple of empty directories to stick — Latest commit e1cd8d8 on Feb 27, 2019

Analysis	Minimal code to reproduce all figures	15 months ago
Data	Trying to get a couple of empty directories to stick	15 months ago
DataOutputs	Trying to get a couple of empty directories to stick	15 months ago
DataProcessing	Minimal code to reproduce all figures	15 months ago
FigureOutputs	Minimal code to reproduce all figures	15 months ago
PCAMV	Minimal code to reproduce all figures	15 months ago
Peripheral	Minimal code to reproduce all figures	15 months ago
Plotting	Minimal code to reproduce all figures	15 months ago
.gitignore	Minimal code to reproduce all figures	15 months ago
README.md	Wording tweak	15 months ago
startup.m	Minimal code to reproduce all figures	15 months ago

README.md

mouseGradients

This repository contains code to reproduce all figures from our paper:

Fulcher, B. D., Murray, J. D., Zerbi, V., & Wang, X.-J. (2019). Multimodal gradients across mouse cortex. PNAS, 201814144.

Figure 3: Correlations across individual cortical layers

```
LaminarPlot()
```

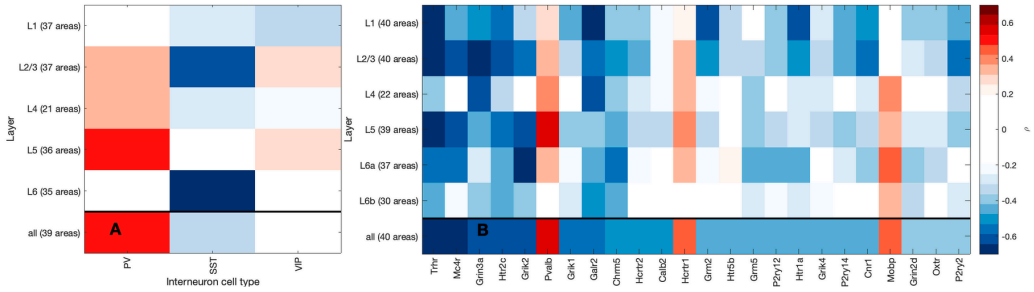
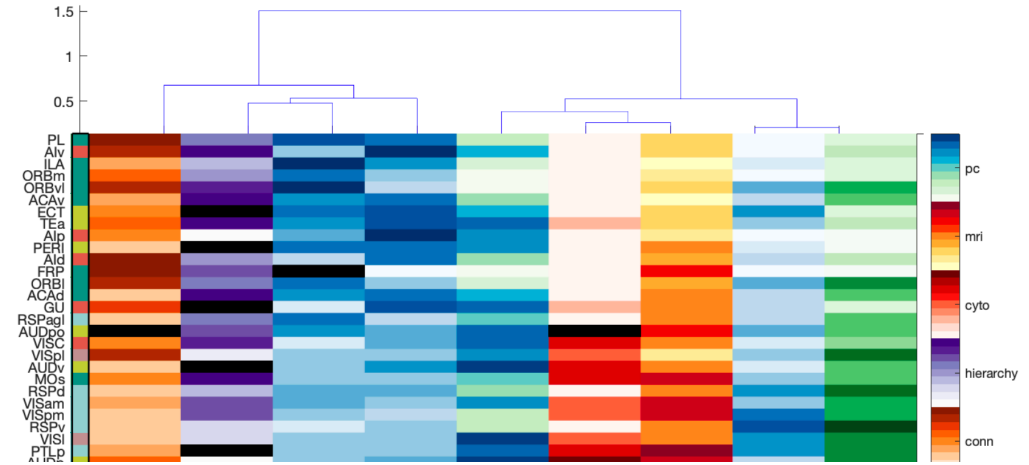


Figure 4: Combination of diverse measurements

```
DiverseMatrixT1T2()
PlotBigDamnMatrix()
```



https://github.com/benfulcher/mouseGradients

Generate a table with key statistics of CFPRs in mouse and human

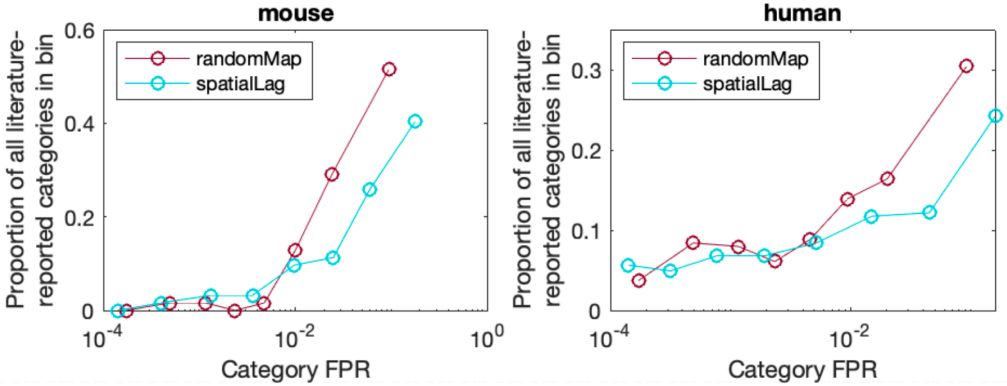
```
FPSRTable();
```

Some key statistics are displayed to the command-line, and outputs full annotated table to `SupplementaryTables/CFPR_Table.csv`.

Investigate the overlap between literature annotations and FPSE as histograms:

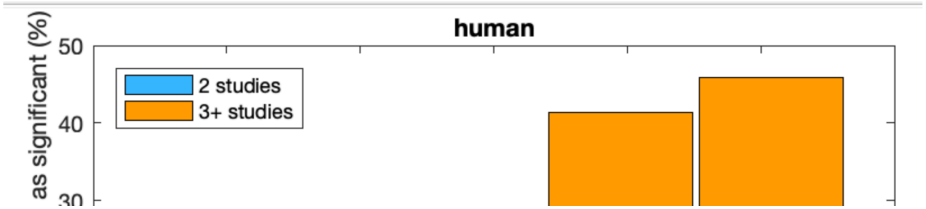
As always, the null phenotype ensemble is defined in the `GiveMeDefaultParams` file.

`propLitCFPR` which looks at how literature-reported categories are distributed across computed levels of CFPR. Outputs figure to `OutputPlots/CFPR_Lit_Together.svg`.



There is also histograms across a linear scale, distinguishing the number of literature analyses flagged across CFPRs:

```
OverlapLitFPSR('mouse',true)
OverlapLitFPSR('human',true)
```



The role of within-category coexpression

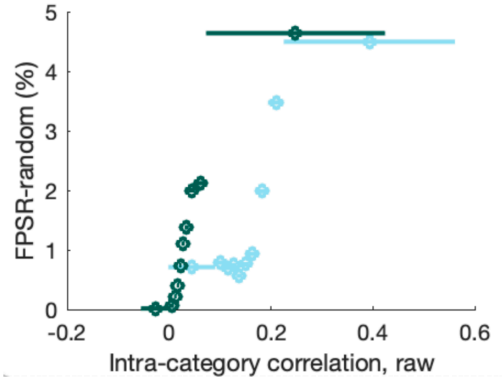
Generate a table characterizing how within-category coexpression varies across GO categories:

```
IntraCorrTable();
```

Outputs to `SupplementaryTables/WithinCategoryCoexp.csv`.

Does intra-category coexpression relate to CFPR?:

```
IntraCorrFPSR();
```



Saves out to `OutputPlots/IntraCorr_CFPR.svg`.

Do categories with spatially autocorrelated genes exhibit an increase in CFPR against spatially autocorrelated ensembles?

Investigate how CFPR is correlated to gene spatial autocorrelation (by category)

```
RelativeFPSRAutoCorr()
```

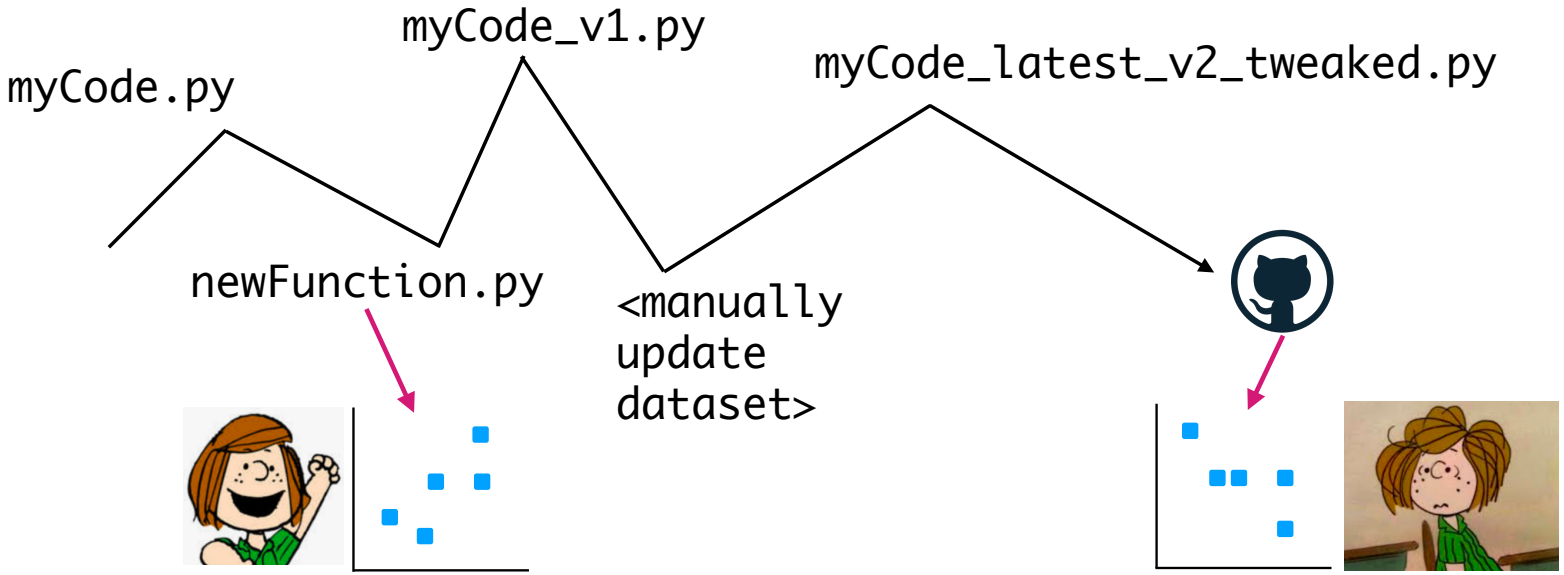
https://github.com/benfulcher/GSEA_FalsePositives

Practices across a Project's Lifespan

The goal of producing a reproducible paper is easier if you work reproducibly across a project's lifespan

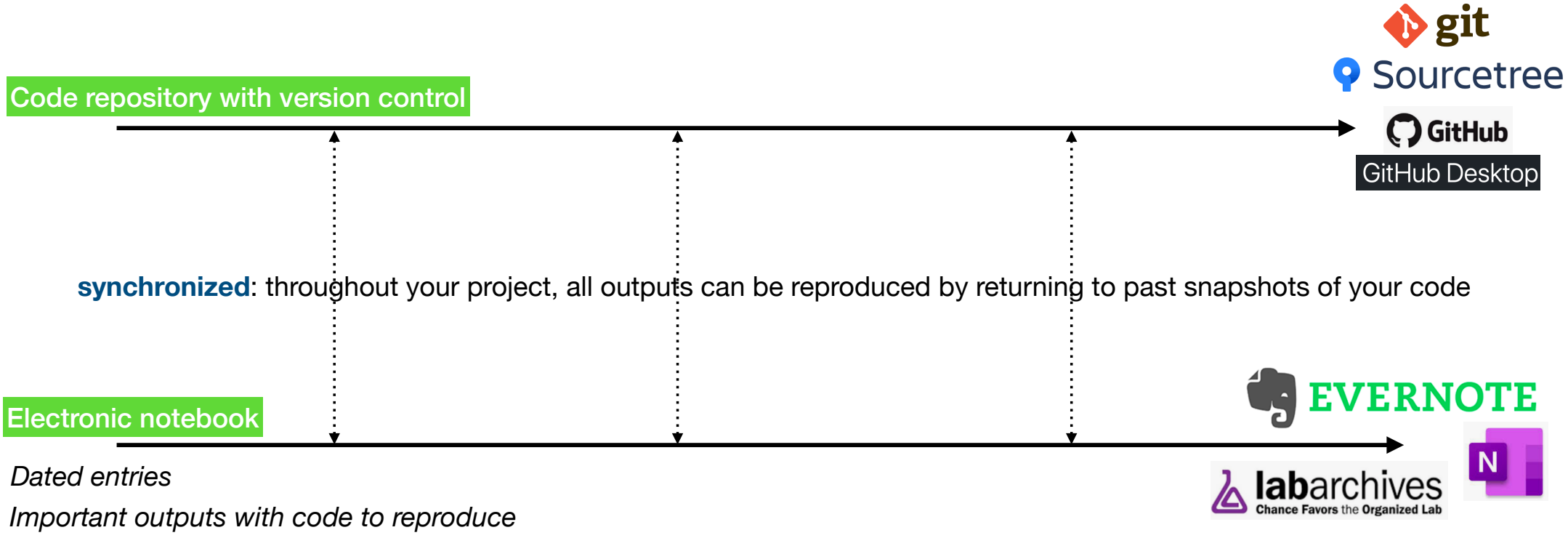
This mindset forces you to work to a much higher standard and pushes your science forward

The scientific process if often a creative one and can have a messy trajectory:



Practices across a Project's Lifespan

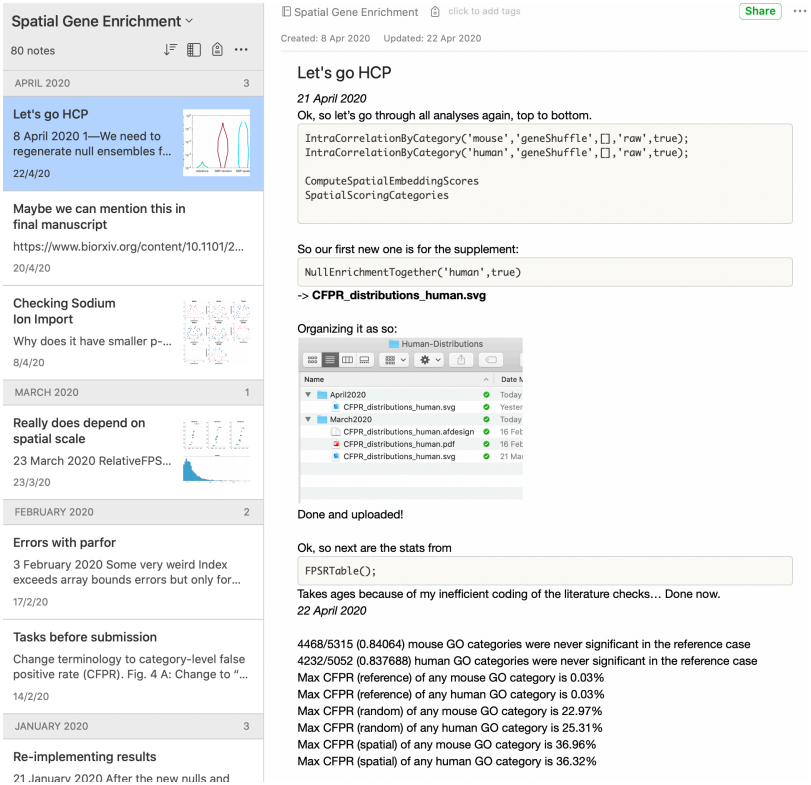
The final goal of producing a reproducible paper is easier if you work reproducibly across a project's lifespan

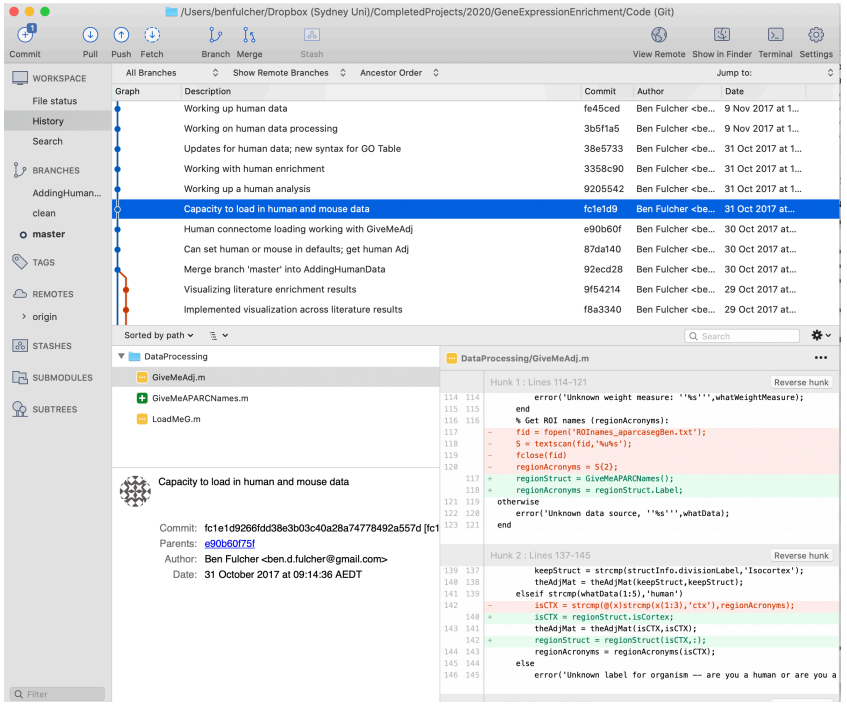


EVERNOTE

git

Sourcetree

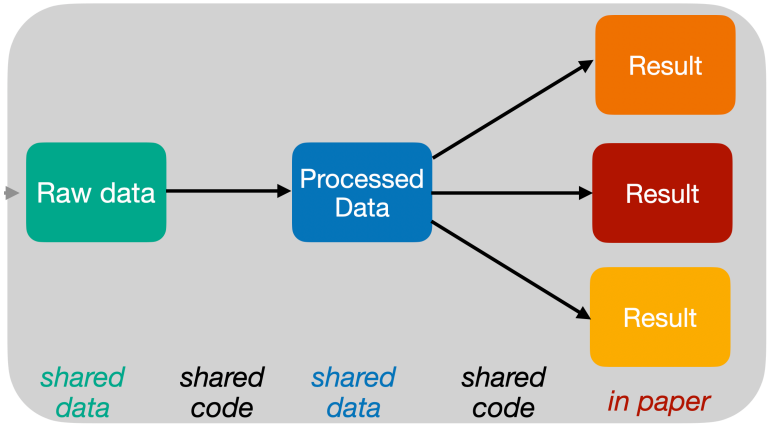




Scientific Process and Outputs

Complete Description of Underlying Methods

Sharing Data



Unsure whether there's a specific data repo?:

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

There are many general and specific scientific data repositories you can use:

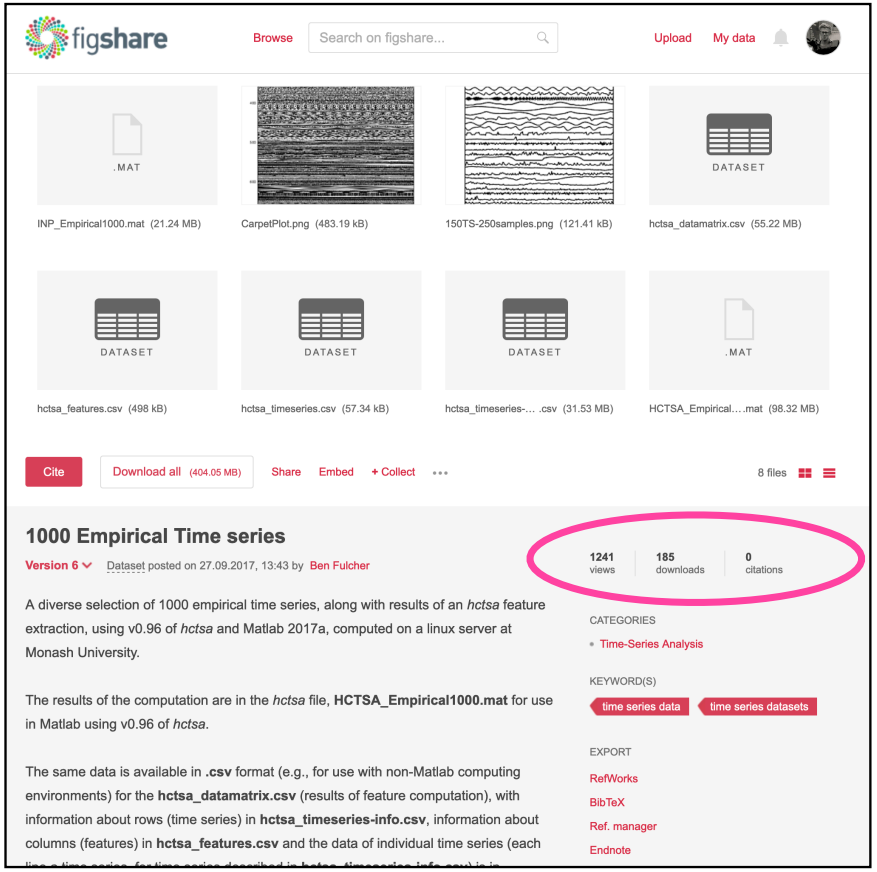
figshare
DRYAD
zenodo

OpenNEURO
A free and open platform for sharing MRI, MEG, EEG, iEEG, and ECoG data

NITRC Neuroimaging Tools & Resources Collaboratory

(most give you a DOI so others can cite your dataset)

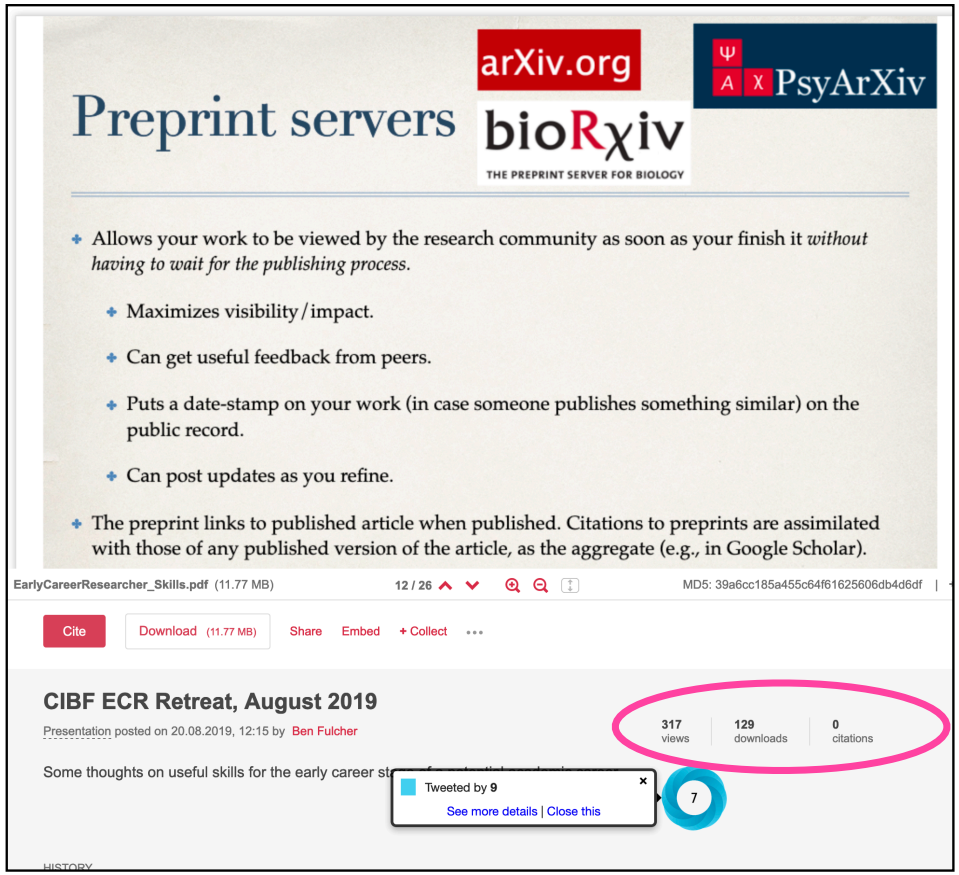
Datasets



https://figshare.com/articles/1000_Empirical_Time_series/5436136

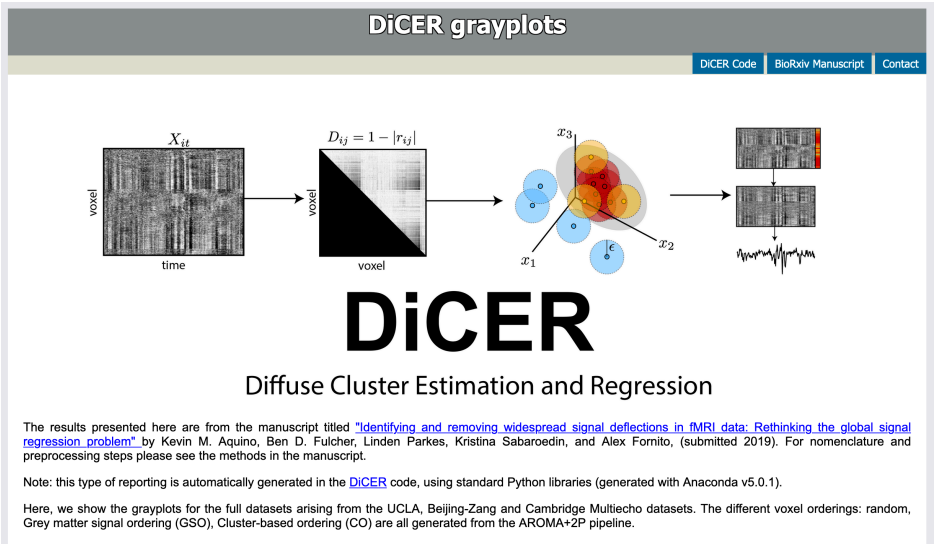
Can track statistics

Slide decks

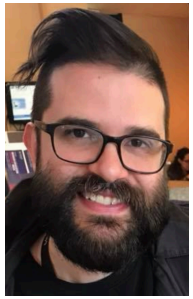


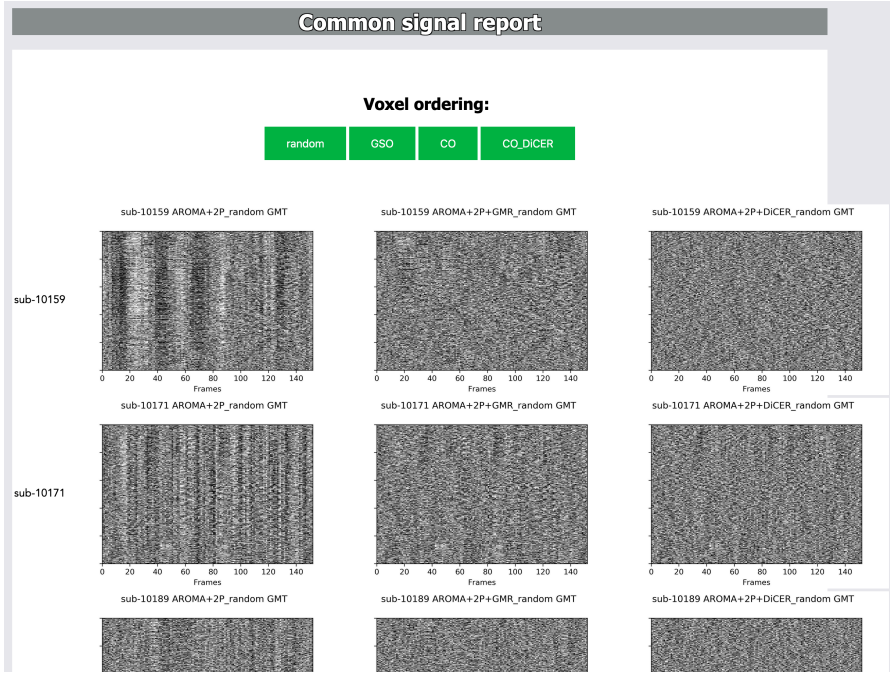
https://figshare.com/articles/CIBF_ECR_Retreat_August_2019/9685415

figshare



https://bmhlab.github.io/DiCER_results/





https://github.com/BMHLab/DiCER

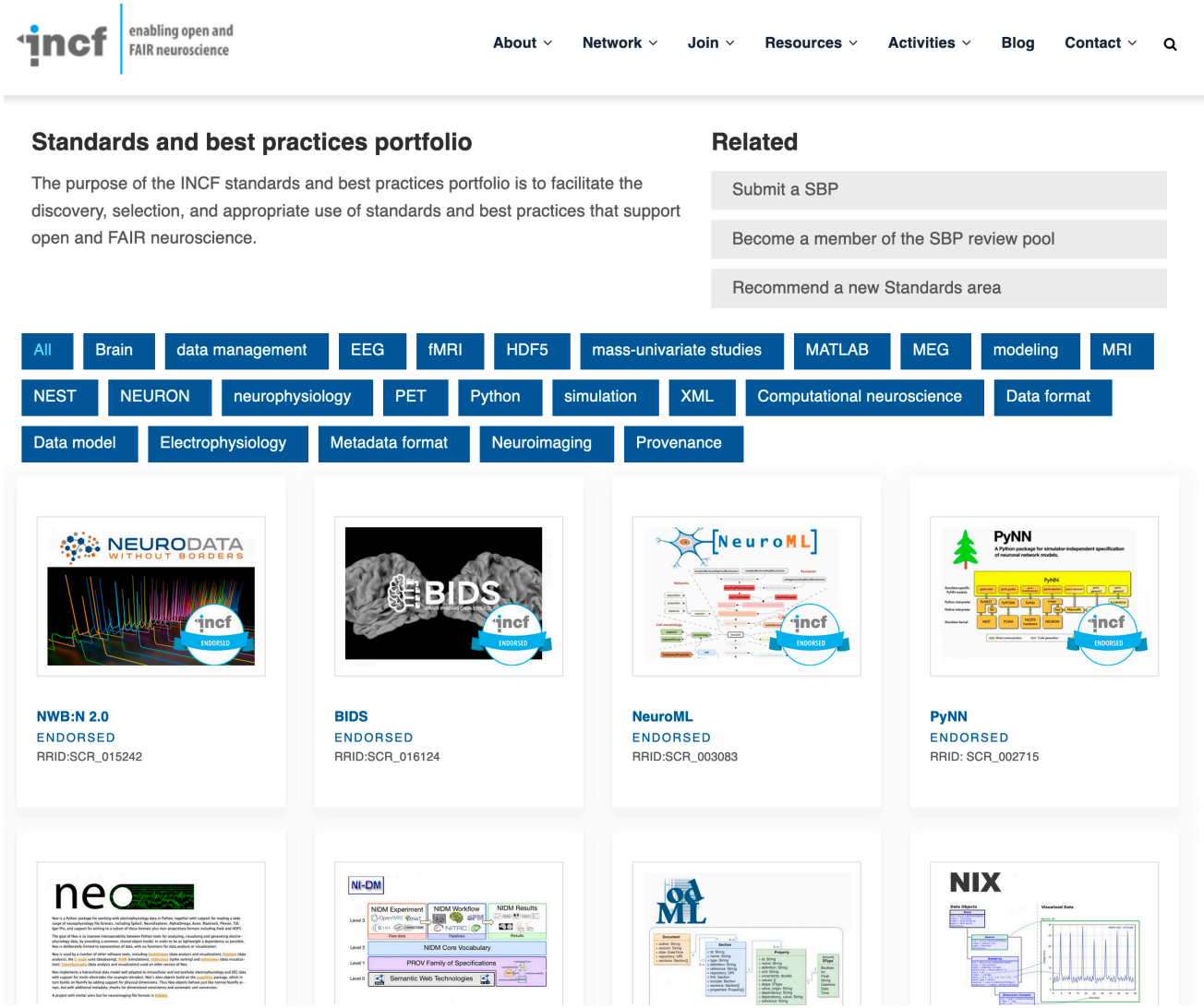
Good Formats

What if everyone shares their data **but** it is poorly annotated or inconsistently formatted?

Is your field in need of standards?:

Recommend a new area for standardization

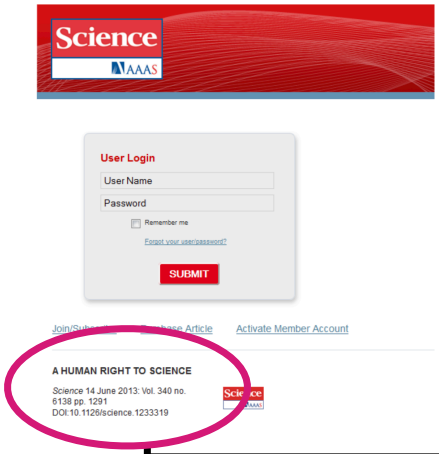
Have you identified an area in need of standardization, or a specific standard or best practice that is missing in neuroscience? Let us know by filling out the form below



https://www.incf.org/resources/sbps

Sharing Your Science

No use making your work reproducible if it's not available in the first place…!



PsyArXiv

bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

Cornell University
arXiv.org

Post-ego science

Open paper

Overleaf AUTHOREA

Open repo

GitHub

Open notebook

EVERNOTE

Closed access only

Preprint (before/ at submission/ publication) [updated at publication]

Open access publication

"Watch me science in real time!"

Sharing Your Science

No use making your work reproducible if it's not available in the first place…!

NHMRC WORKING TO BUILD A HEALTHY AUSTRALIA

Dissemination of Research Findings

NHMRC Policy	ANY NHMRC-supported, peer-reviewed journal article published after the 1st of July 2012 must be made freely available via an open access repository or journal website by the CIA within twelve months from the date of publication.
.	The metadata must be provided to your Institutional Repository (IR) immediately after the publication date.
.	For further information, see the revised NHMRC policy on the Dissemination of Research Findings: http://www.nhmrc.gov.au/grants/policy/dissemination-research-findings
Have You Deposited The Metadata Of This Publication To An IR?	No
Is This Article Now Freely Available Via An Open Access Repository?	No (eg IR, PubMedCentral, arXiv, journal website)

Sharing Your Science

You can also bypass the for-profit journal system altogether

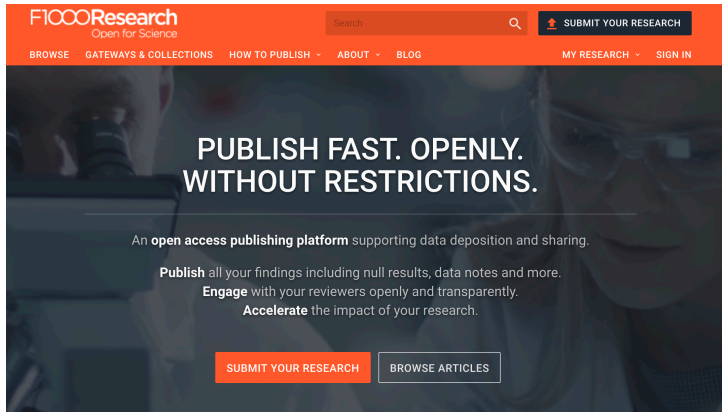
Overlay journals: community run with open peer-review, and either free or near-free (~$20)

(we are not post-ego yet): Check your uni acknowledges these journals (e.g., some may not be Scopus indexed)

PsyArXiv

bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

Cornell University
arXiv.org



https://f1000research.com/

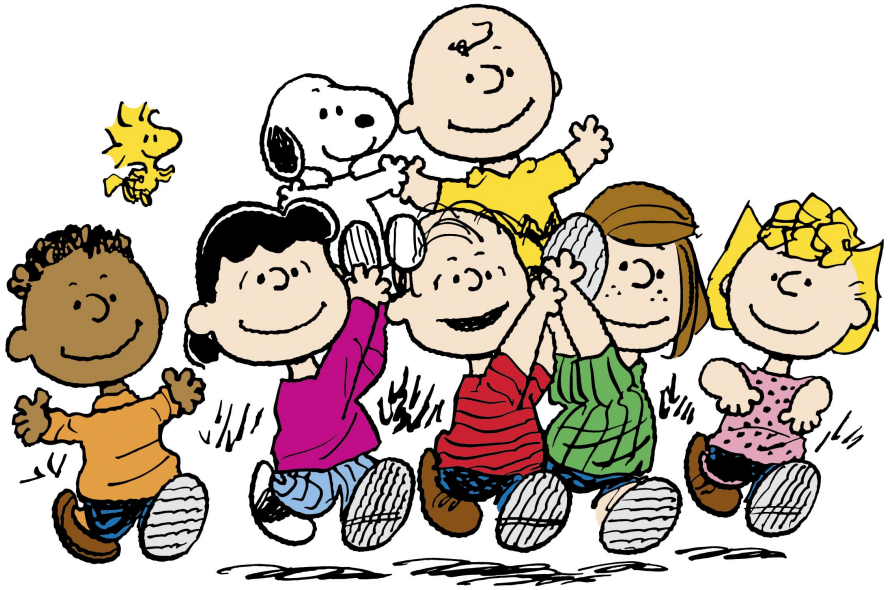
Neurons, Behavior, Data analysis, and Theory

https://nbdt.scholasticahq.com/

Summary

Open research practices are unambiguously good for science and usually also good for your career (e.g., more exposure, greater impact)

It can be hard to change habits, and there is a learning curve (esp. for non-computational scientists) but go at your own pace, support others, and we'll incrementally build a better science



SCIENTIFIC STANDARDS

Promoting an open research culture

Author guidelines for journals could help to promote transparency, openness, and reproducibility

By B. A. Nosek,* G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J. Freese, R. Glennerster, D. Goroff, D. P. Green, B. Hesse, M. Humphreys, J. Ishiyama, D. Karlan, A. Kraut, A. Lupia, P. Mabry, T. Madon, N. Malhotra, E. Mayo-Wilson, M. McNutt, E. Miguel, E. Levy Paluck, U. Simonsohn, C. Soderberg, B. A. Spellman, J. Turitto, G. VandenBos, S. Vazire, E. J. Wagenmakers, R. Wilson, T. Yarkoni

ben.fulcher@sydney.edu.au
@bendfulcher
benfulcher
www.benfulcher.com

THE UNIVERSITY OF SYDNEY