

# **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)

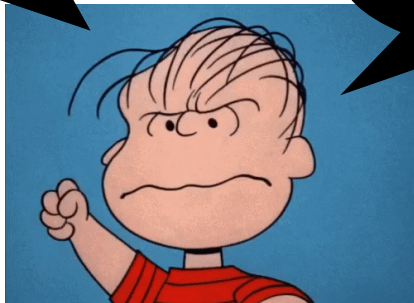
**Any little step** you make in the direction of open science is progress

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

*"I want to reduce suffering of animals"*



How dare you  
eat honey!



A leather belt?  
MURDERER!

*"I want to do open science"*



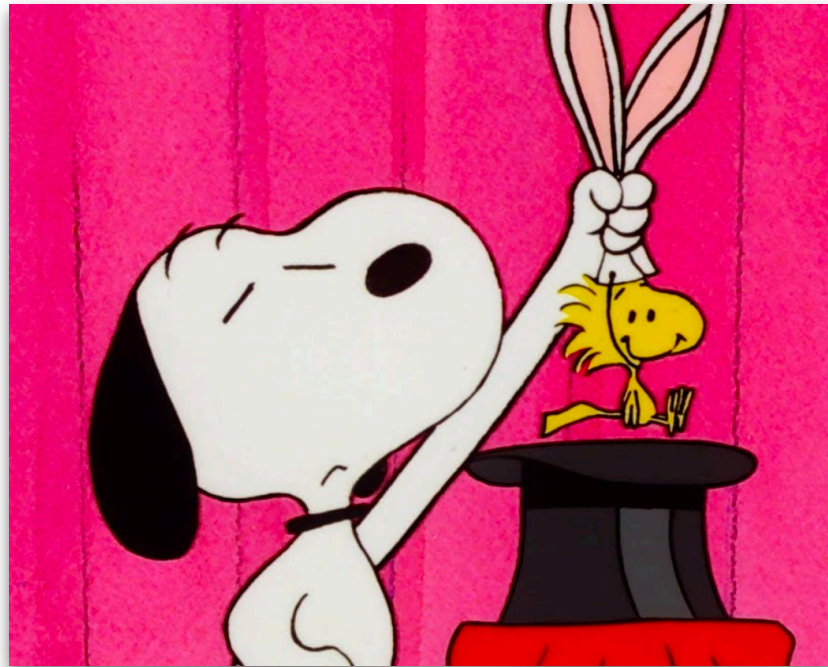
Amateur! You  
didn't use  
docker?



You use Matlab?  
You must HATE  
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.”



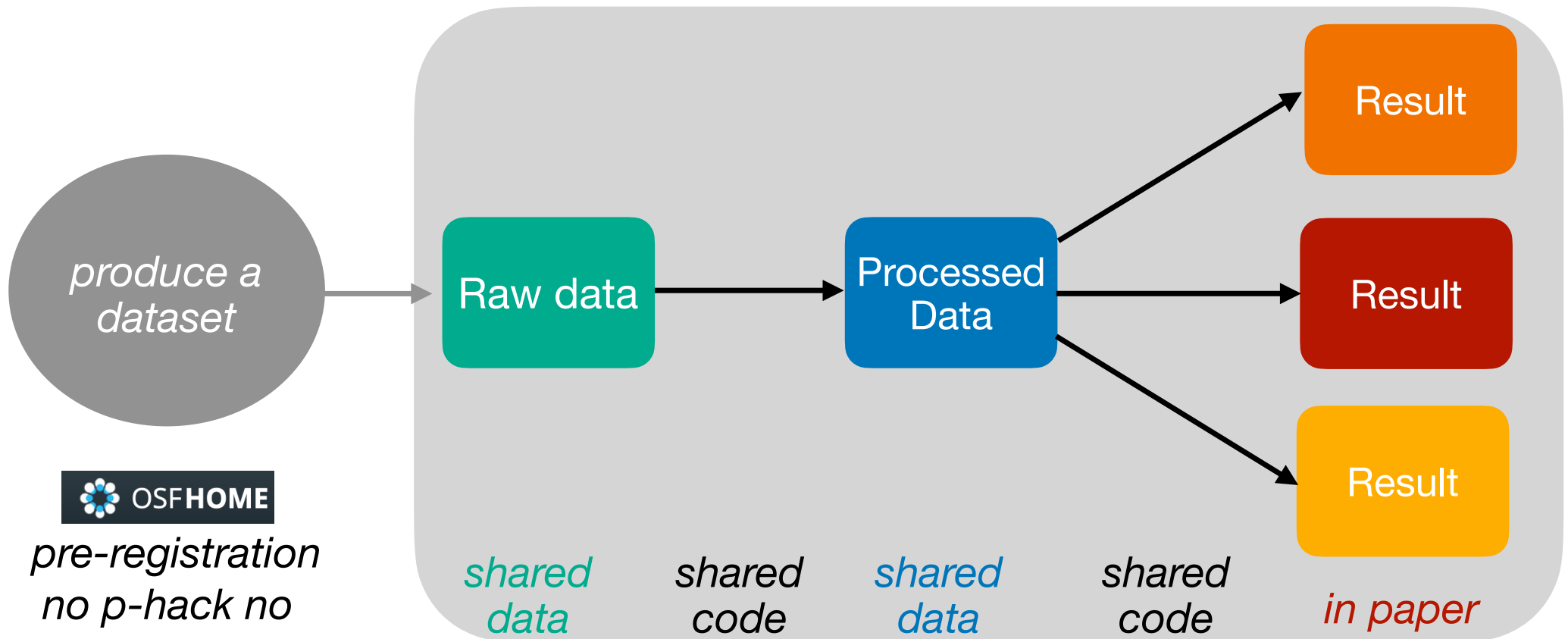
~Donohoe

<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.”

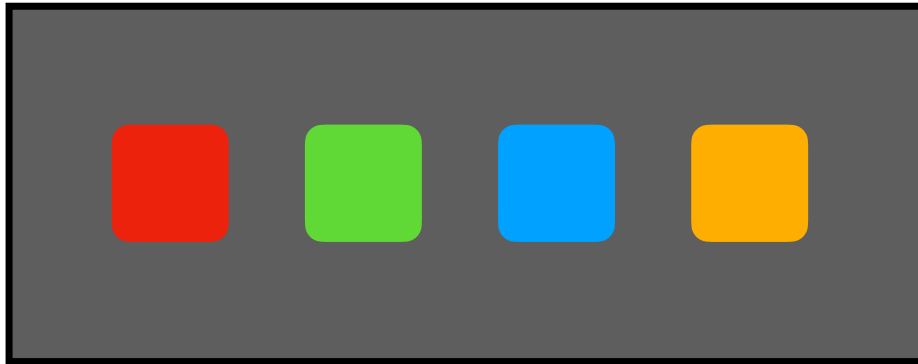


# Level 0

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"*

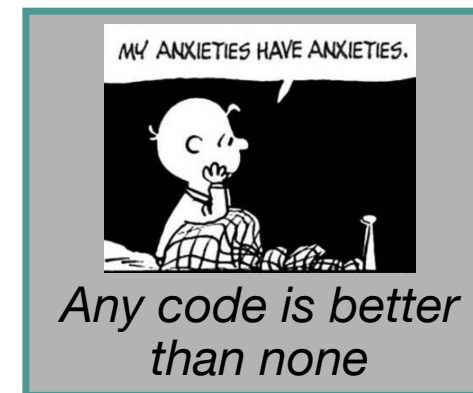


# Level 1

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
```



# 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



Zenodo now supports DOI  
versioning!  
[Read more](#) about it, in our newest blog  
post.



Using GitHub?

Just [Log in](#) with your GitHub account and  
[click here](#) to start preserving your  
repositories.



# 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

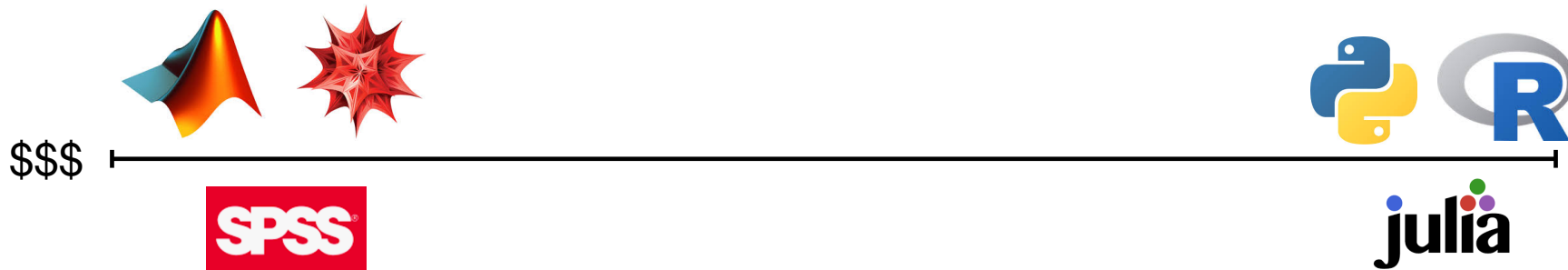


<https://www.docker.com/>

# 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

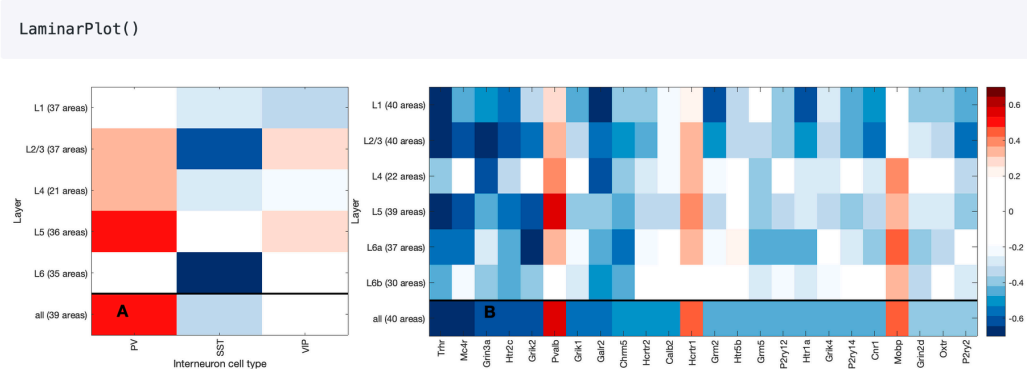
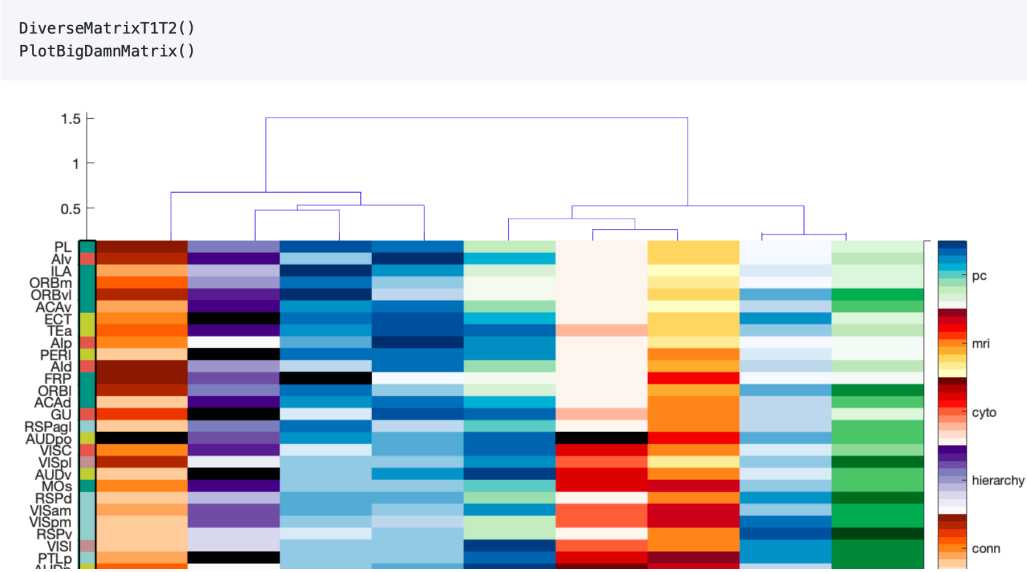


Figure 4: Combination of diverse measurements



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

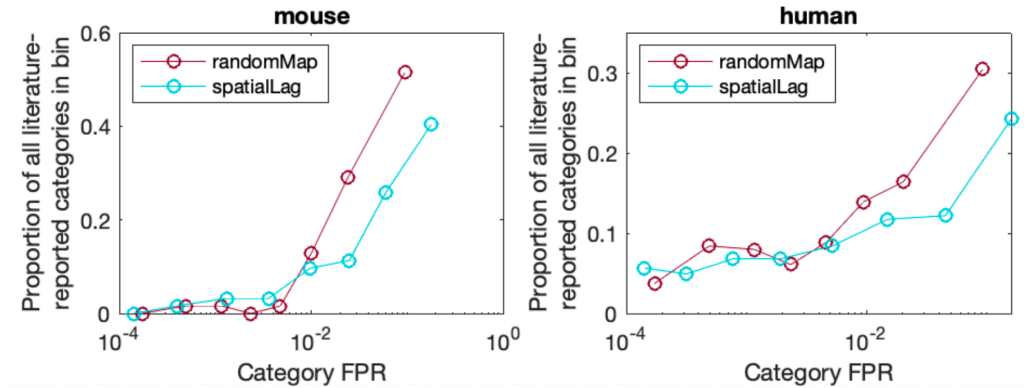
```
FPSRTable();
```

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

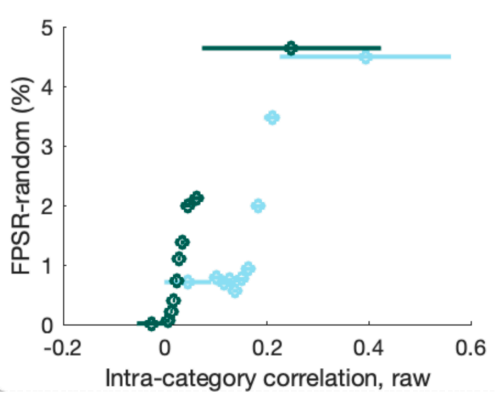
Generate a table characterizing how within-category coexpression varies across GO categories:

```
IntraCorrTable();
```

Outputs to `SupplementaryTables/WithinCategoryCoexp.csv`.

Does intra-category coexpression relate to CFPR?:

```
IntraCorrFPSR();
```

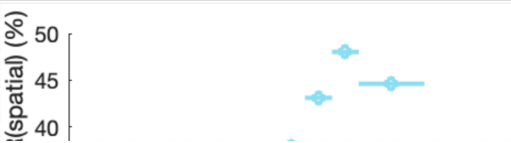


Saves out to `OutputPlots/IntraCorr_CFPR.svg`.

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()
```

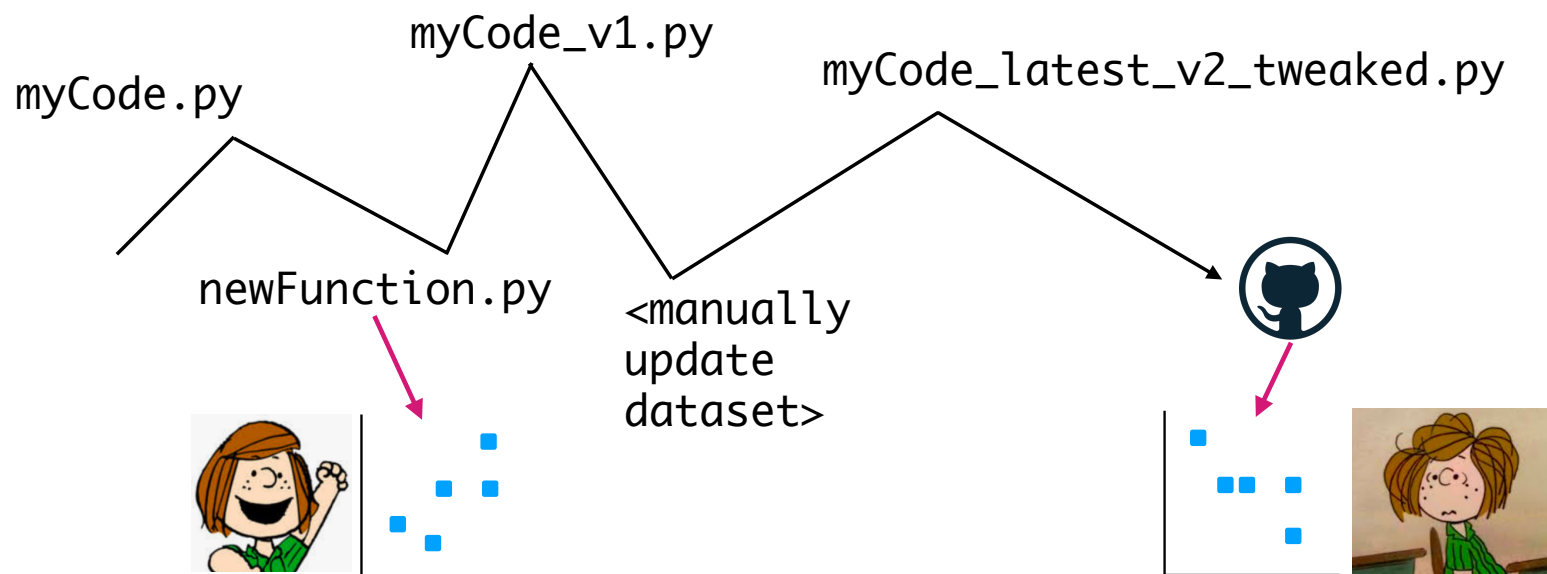


# Practices across a Project's Lifespan

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

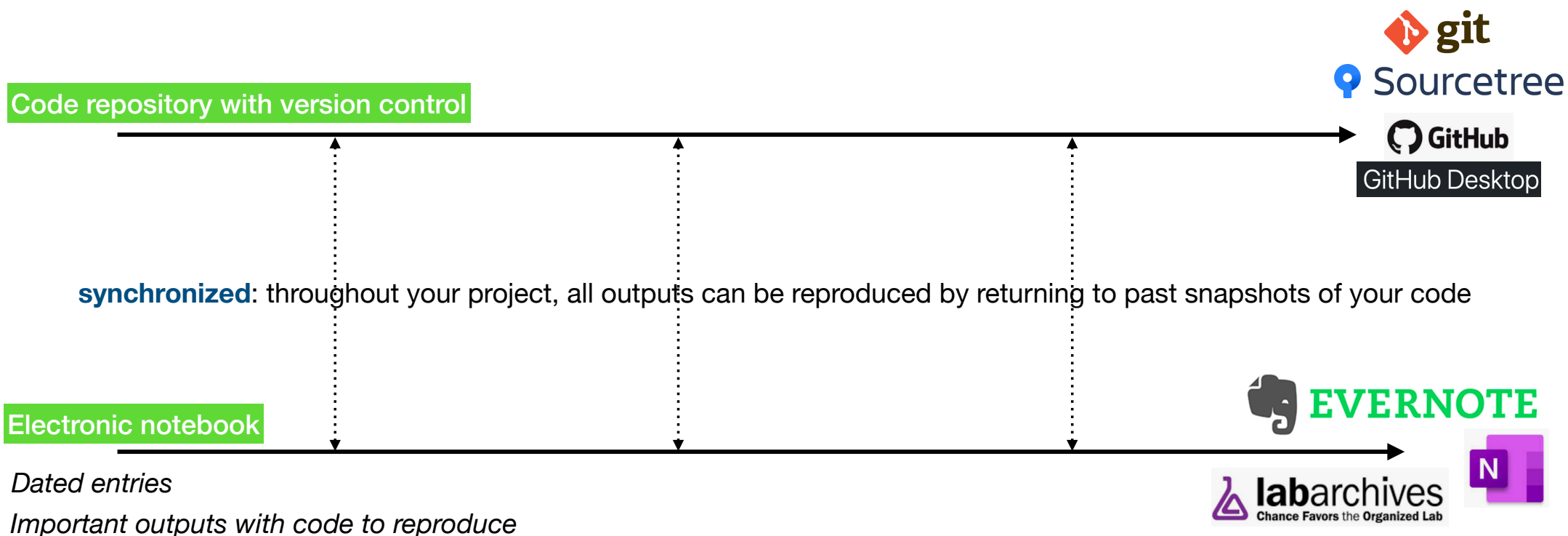
This mindset forces you to work to a much higher standard and pushes your science forward

The scientific process is 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





Spatial Gene Enrichment

80 notes

APRIL 2020 3

Let's go HCP

8 April 2020 1—We need to regenerate null ensembles f...

22/4/20

Maybe we can mention this in final manuscript

<https://www.biorxiv.org/content/10.1101/2...>

20/4/20

Checking Sodium Ion Import

Why does it have smaller p-...

8/4/20

MARCH 2020 1

Really does depend on spatial scale

23 March 2020 RelativeFPS...

23/3/20

FEBRUARY 2020 2

Errors with parfor

3 February 2020 Some very weird Index exceeds array bounds errors but only for...

17/2/20

Tasks before submission

Change terminology to category-level false positive rate (CFPR). Fig. 4 A: Change to "...

14/2/20

JANUARY 2020 3

Re-implementing results

21 January 2020 After the new nulls and

Spatial Gene Enrichment

Created: 8 Apr 2020 Updated: 22 Apr 2020

Share

Let's go HCP

21 April 2020

Ok, so let's go through all analyses again, top to bottom.

```
IntraCorrelationByCategory('mouse','geneShuffle',[],'raw',true);  
IntraCorrelationByCategory('human','geneShuffle',[],'raw',true);  
  
ComputeSpatialEmbeddingScores  
SpatialScoringCategories
```

So our first new one is for the supplement:

```
NullEnrichmentTogether('human',true)
```

-> CFPR\_distributions\_human.svg

Organizing it as so:

Done and uploaded!

Ok, so next are the stats from

```
FPSRTTable();
```

Takes ages because of my inefficient coding of the literature checks... Done now.

22 April 2020

4488/5315 (0.84064) mouse GO categories were never significant in the reference case

4232/5052 (0.837688) human GO categories were never significant in the reference case

Max CFPR (reference) of any mouse GO category is 0.03%

Max CFPR (reference) of any human GO category is 0.03%

Max CFPR (random) of any mouse GO category is 22.97%

Max CFPR (random) of any human GO category is 25.31%

Max CFPR (spatial) of any mouse GO category is 36.96%

Max CFPR (spatial) of any human GO category is 36.32%

Commit Pull Push Fetch Branch Merge Stash

View Remote Show in Finder Terminal Settings

Commit Author Date

Working up human data fe45ced Ben Fulcher <ben.d.fulcher@gmail.com> 9 Nov 2017 at 1...

Working on human data processing 3b5f1a5 Ben Fulcher <ben.d.fulcher@gmail.com> 9 Nov 2017 at 1...

Updates for human data; new syntax for GO Table 38e6733 Ben Fulcher <ben.d.fulcher@gmail.com> 31 Oct 2017 at 1...

Working with human enrichment 3358c90 Ben Fulcher <ben.d.fulcher@gmail.com> 31 Oct 2017 at 1...

Working up a human analysis 9205542 Ben Fulcher <ben.d.fulcher@gmail.com> 31 Oct 2017 at 1...

Capacity to load in human and mouse data fc1e1d9 Ben Fulcher <ben.d.fulcher@gmail.com> 31 Oct 2017 at 1...

Human connectome loading working with GiveMeAdj e90b60f Ben Fulcher <ben.d.fulcher@gmail.com> 30 Oct 2017 at 1...

Can set human or mouse in defaults; get human Adj 87da140 Ben Fulcher <ben.d.fulcher@gmail.com> 30 Oct 2017 at 1...

Merge branch 'master' into AddingHumanData 92ecd28 Ben Fulcher <ben.d.fulcher@gmail.com> 30 Oct 2017 at 1...

Visualizing literature enrichment results 9f54214 Ben Fulcher <ben.d.fulcher@gmail.com> 29 Oct 2017 at 1...

Implemented visualization across literature results f8a3340 Ben Fulcher <ben.d.fulcher@gmail.com> 29 Oct 2017 at 1...

Sorted by path

DataProcessing/GiveMeAdj.m

Hunk 1: Lines 114-121

```
error('Unknown weight measure: '%s'',whatWeightMeasure);  
end  
% Get ROI names (regionAcronyms):  
fid = fopen('ROINames_aparcasegen.txt');  
S = textscan(fid,'%u%');  
fclose(fid);  
regionAcronyms = S(2);  
regionStruct = GiveMeAPARNames();  
regionAcronyms = regionStruct.Label;  
otherwise  
error('Unknown data source, '%s'',whatData);  
end
```

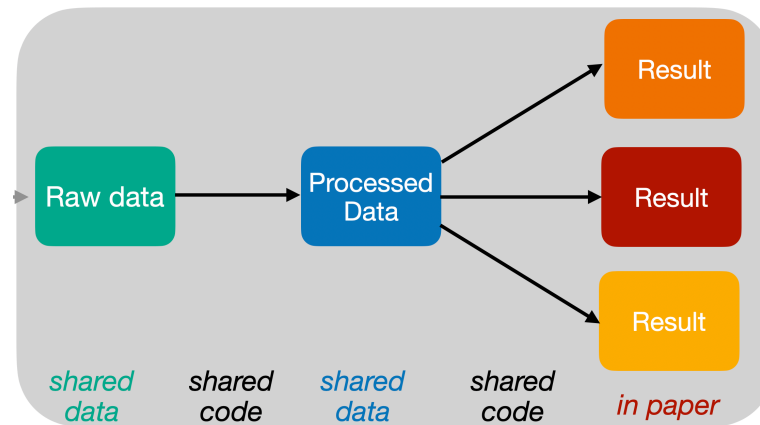
Hunk 2: Lines 137-145

```
keepStruct = struct(structInfo.divisionLabel,'Isocortex');  
theadMat = theAdjMat(keepStruct,keepStruct);  
elseif strcmp(whatData(1:15),'human')  
isCTX = strcmp(@(isCTX)isCTX(1:3),'ctx'),regionAcronyms);  
theadMat = theAdjMat(isCTX,isCTX);  
regionStruct = regionStruct(isCTX,:);  
regionAcronyms = regionAcronyms(isCTX);  
else  
error('Unknown label for organism — are you a human or are you a
```

Scientific Process and Outputs

Complete Description of Underlying Methods

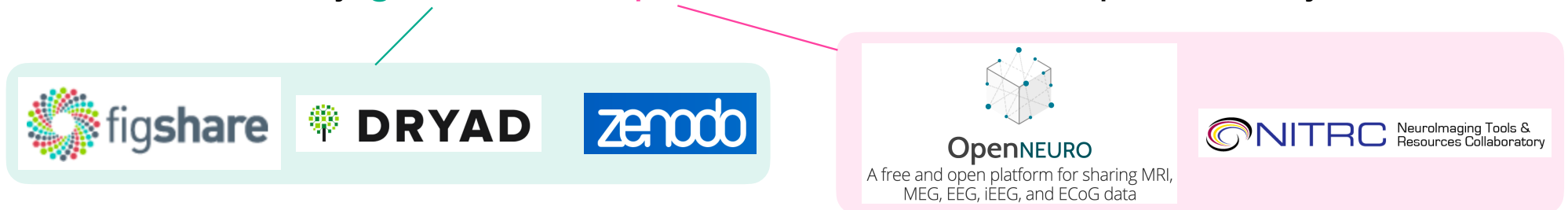
# Sharing Data



Unsure whether there's a specific data repo?:



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



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

# Datasets

The screenshot shows a Figshare dataset page. At the top, there's a search bar and navigation links. Below, a grid of file thumbnails is displayed, including .MAT files and CSV datasets. A pink oval highlights the statistics section on the right, which shows 1241 views, 185 downloads, and 0 citations. The main description text is visible on the left, detailing the dataset's origin and format.



Can track statistics

# Slide decks

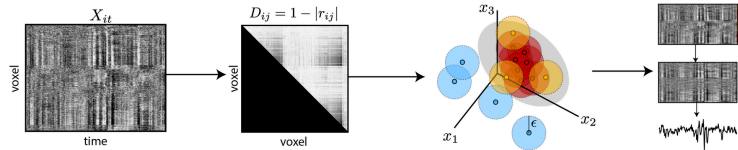
The screenshot shows a Figshare preprint server page. At the top, there are logos for arXiv.org, bioRxiv, and PsyArXiv. The main title is 'Preprint servers'. Below, a list of benefits for preprints is shown. A pink oval highlights the statistics section on the right, which shows 317 views, 129 downloads, and 0 citations. The main description text is visible on the left, detailing the presentation's content.

[https://figshare.com/articles/CIBF\\_ECR\\_Retreat\\_August\\_2019/9685415](https://figshare.com/articles/CIBF_ECR_Retreat_August_2019/9685415)





## DiCER grayplots

[DiCER Code](#)[BioRxiv Manuscript](#)[Contact](#)

# DiCER

## Diffuse Cluster Estimation and Regression

The results presented here are from the manuscript titled "[Identifying and removing widespread signal deflections in fMRI data: Rethinking the global signal regression problem](#)" by Kevin M. Aquino, Ben D. Fulcher, Linden Parkes, Kristina Sabaroein, and Alex Fornito, (submitted 2019). For nomenclature and preprocessing steps please see the methods in the manuscript.

Note: this type of reporting is automatically generated in the [DiCER](#) code, using standard Python libraries (generated with Anaconda v5.0.1).

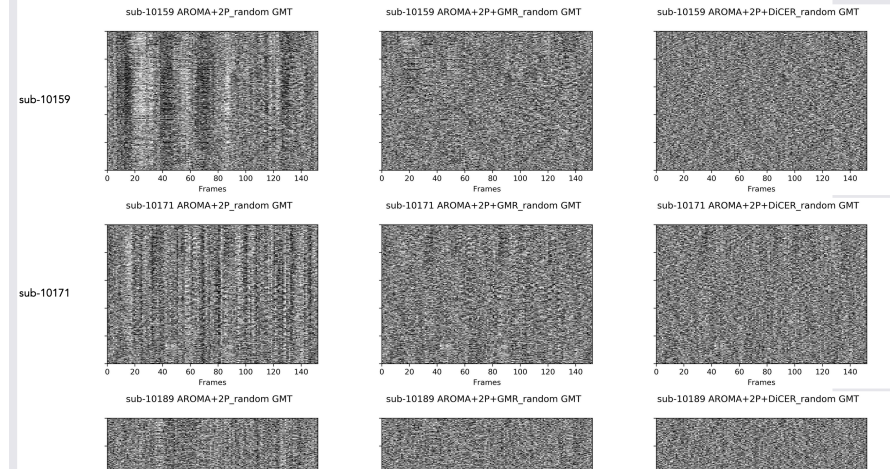
Here, we show the grayplots for the full datasets arising from the UCLA, Beijing-Zang and Cambridge Multiecho datasets. The different voxel orderings: random, Grey matter signal ordering (GSO), Cluster-based ordering (CO) are all generated from the AROMA+2P pipeline.

[https://bmhlab.github.io/DiCER\\_results/](https://bmhlab.github.io/DiCER_results/)



## Common signal report

### Voxel ordering:

[random](#)[GSO](#)[CO](#)[CO\\_DICER](#)

<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

## Standards and best practices portfolio

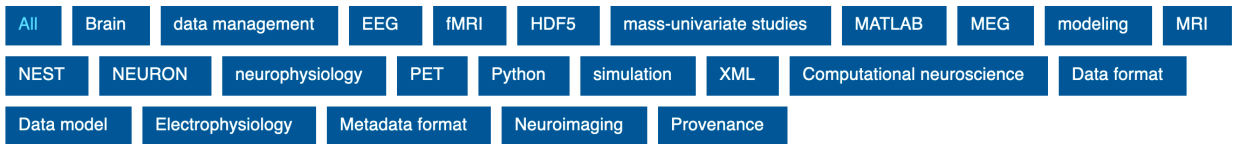
The purpose of the INCF standards and best practices portfolio is to facilitate the discovery, selection, and appropriate use of standards and best practices that support open and FAIR neuroscience.

## Related

[Submit a SBP](#)

[Become a member of the SBP review pool](#)

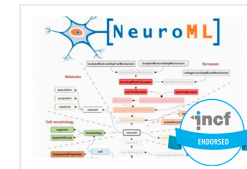
[Recommend a new Standards area](#)



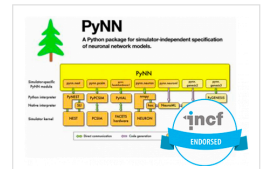
**NWB:N 2.0**  
ENDORSED  
RRID:SCR\_015242



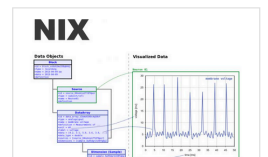
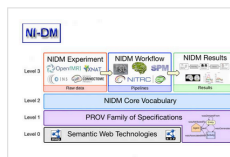
**BIDS**  
ENDORSED  
RRID:SCR\_016124



**NeuroML**  
ENDORSED  
RRID:SCR\_003083



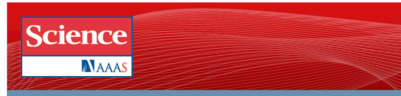
**PyNN**  
ENDORSED  
RRID:SCR\_002715



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

# Sharing Your Science

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



User Login

User Name

Password

☐ Remember me

[Forgot your username/password?](#)

SUBMIT

[Join/Subscribe](#) [Share Article](#) [Activate Member Account](#)

## A HUMAN RIGHT TO SCIENCE

Science 14 June 2013; Vol. 340 no. 6138 pp. 1291  
DOI:10.1126/science.1233319



Post-ego  
science

Open paper



Open repo



Open notebook



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

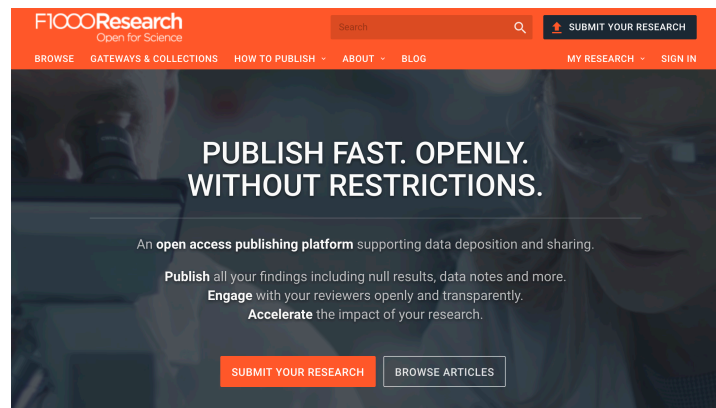
( eg IR, PubMedCentral, arXiv, journal website )

# 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)



<https://f1000research.com/>

Neurons, Behavior, Data analysis, and  
Theory

<https://nbd.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](mailto:ben.fulcher@sydney.edu.au)

 [@bendfulcher](https://twitter.com/bendfulcher)

 [benfulcher](https://github.com/benfulcher)

[www.benfulcher.com](http://www.benfulcher.com)



THE UNIVERSITY OF  
SYDNEY