

Digital badge

- Reproducible research

Brendan Palmer,

Clinical Research Facility - Cork &

School of Public Health

 @B_A_Palmer



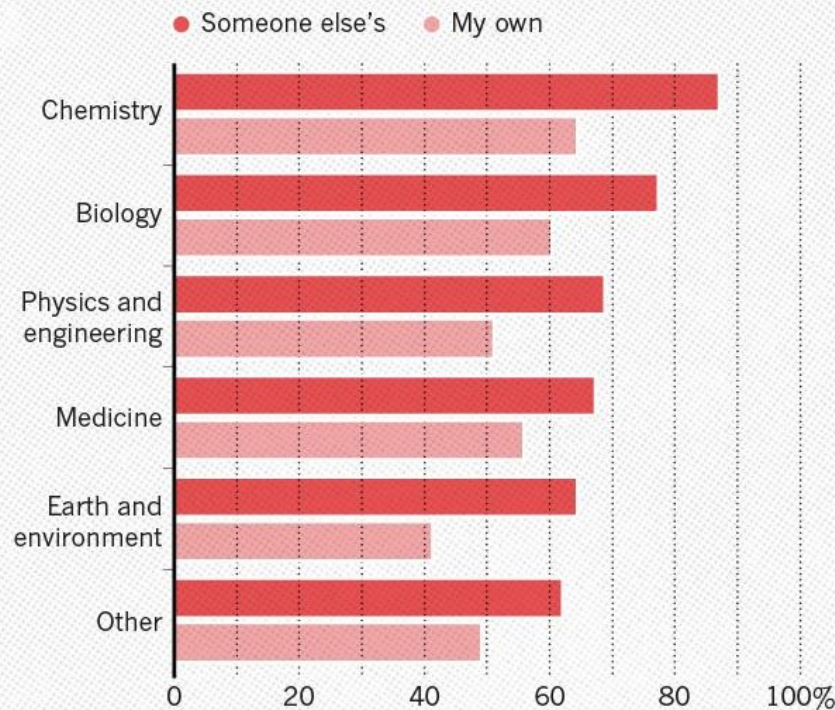
1,500 scientists lift the lid on reproducibility

Survey sheds light on the 'crisis' rocking research.

[Monya Baker](#)

HAVE YOU FAILED TO REPRODUCE AN EXPERIMENT?

Most scientists have experienced failure to reproduce results.



COMMENTARY

Scientists behaving badly

To protect the integrity of science, we must look beyond falsification, fabrication and plagiarism, to a wider range of questionable research practices, argue **Brian C. Martinson**, **Melissa S. Anderson** and **Raymond de Vries**.

Table 1 | Percentage of scientists who say that they engaged in the behaviour listed within the previous three years (n = 3,247)

Top ten behaviours	All	Mid-career	Early-career
1. Falsifying or 'cooking' research data	0.3	0.2	0.5
2. Ignoring major aspects of human-subject requirements	0.3	0.3	0.4
3. Not properly disclosing involvement in firms whose products are based on one's own research	0.3	0.4	0.3
4. Relationships with students, research subjects or clients that may be interpreted as questionable	1.4	1.3	1.4
5. Using another's ideas without obtaining permission or giving due credit	1.4	1.7	1.0
6. Unauthorized use of confidential information in connection with one's own research	1.7	2.4	0.8 ***
7. Failing to present data that contradict one's own previous research	6.0	6.5	5.3
8. Circumventing certain minor aspects of human-subject requirements	7.6	9.0	6.0 **
9. Overlooking others' use of flawed data or questionable interpretation of data	12.5	12.2	12.8
10. Changing the design, methodology or results of a study in response to pressure from a funding source	15.5	20.6	9.5 ***
Other behaviours			
11. Publishing the same data or results in two or more publications	4.7	5.9	3.4 **
12. Inappropriately assigning authorship credit	10.0	12.3	7.4 ***
13. Withholding details of methodology or results in papers or proposals	10.8	12.4	8.9 **
14. Using inadequate or inappropriate research designs	13.5	14.6	12.2
15. Dropping observations or data points from analyses based on a gut feeling that they were inaccurate	15.3	14.3	16.5
16. Inadequate record keeping related to research projects	27.5	27.7	27.3

Who benefits most from reproducibility?



Casey Greene

@GreeneScientist

Follow



Reproducibility is important because the you of 3 months ago is terrible at answering email! - [@tracykteal](#) at [#2016dssummit](#)

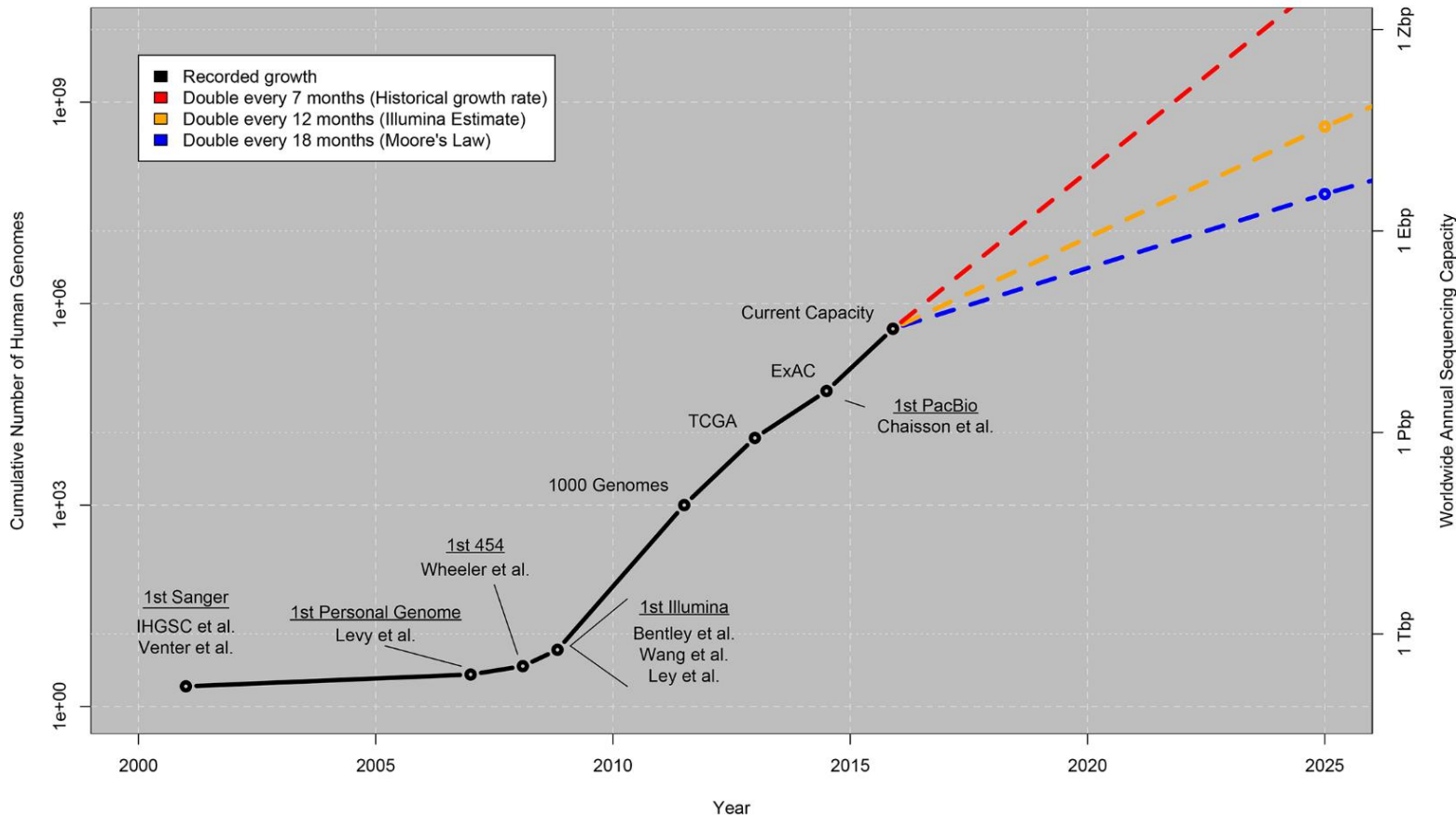
1:17 PM - 26 Oct 2016 from [Manhattan, NY](#)

Where to begin...



The challenge

Growth of DNA Sequencing



Astrophiz
@Astrophiz

Follow

Congratulations to Dr Katie Bouman!
This is the woman who created the algorithm to crunch the 5 petabytes of data from 500 kg of hard drives from 8 radio telescopes to make the first image of the [#EHTBlackHole](#) [#BlackHole](#)



2:55 PM - 10 Apr 2019

Fundamental problem



I'm not in the office at the moment. Send any work to be translated

Beware of default settings

Ziemann et al. *Genome Biology* (2016) 17:177
DOI 10.1186/s13059-016-1044-7

Genome Biology

COMMENT

Open Access



Gene name errors are widespread in the scientific literature

Mark Ziemann¹, Yotam Eren^{1,2} and Assam El-Osta^{1,3*}

Abstract

The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

frequently reused. Our aim here is to raise awareness of the problem.

We downloaded and screened supplementary files from 18 journals published between 2005 and 2015 using a suite of shell scripts. Excel files (.xls and .xlsx suffixes) were converted to tabular separated files (tsv) with ssconvert (v1.12.9). Each sheet within the Excel file was converted to a separate tsv file. Each column of data in the tsv file was screened for the presence of gene sym-

NCBI LocusLink

Search LocusLink Display Brief Organism: All

Query: Go Clear

View Hs NEDD5 One of 1 Loci Save All Loci

Click to Display tRNA-Genomic Alignments (spanning 38716 bps)

PUB OMIM REVIEW UNIGENE MAP VAR HOMOL GDB

ef UCSC

Homo sapiens Official Gene Symbol and Name (HGNC)

NEDD5: neural precursor cell expressed, developmentally down-regulated 5

LocusID: 4735

Overview [Submit GeneRIF](#) ?

Locus Type: gene with protein product, function known or inferred

Product: neural precursor cell expressed, developmentally down-regulated 5

Alternate Symbols: DIFF6, SEPT2, hNedd5, KIAA0158

Relationships ?

Mouse Homology Maps:

NCBI vs. MGD	1 cM	2-Sep	Hs Mm
UCSC vs. MGD	1 cM	Sept2	Hs Mm
UCSC vs. Hudson et al.	1 1319.34 cR	AW208991	Hs Mm

Map Information ?

Beware of default settings



The Reinhart-Rogoff error – or how not to Excel at economics

April 22, 2013 9.40pm BST

Data and computer code should be made publicly available at an early stage – or else ... esarastudillo

✉ Email

🐦 Twitter

📘 Facebook

🌐 LinkedIn

🖨️ Print

88

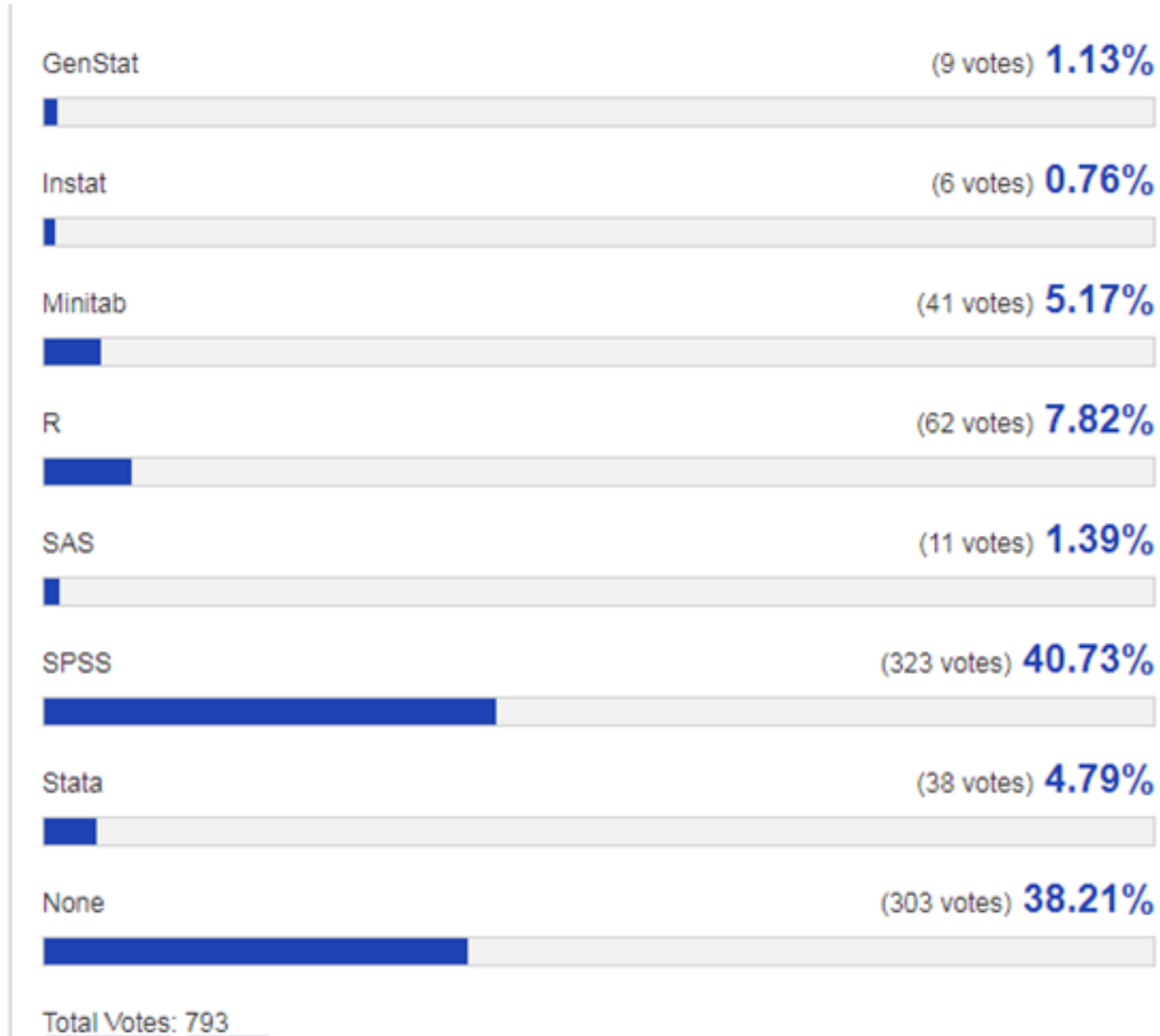
453

Last week we learned a famous [2010 academic paper](#), relied on by political big-hitters to bolster arguments for austerity cuts, contained significant errors; and that those errors came down to misuse of an Excel spreadsheet.

Sadly, these are not the first mistakes of this size and nature when handling data. So what on Earth went wrong, and can we fix it?

Harvard's [Carmen Reinhart](#) and [Kenneth Rogoff](#) are two of the most respected and influential academic economists active today.

Where are we all coming from?



Putting the pieces together

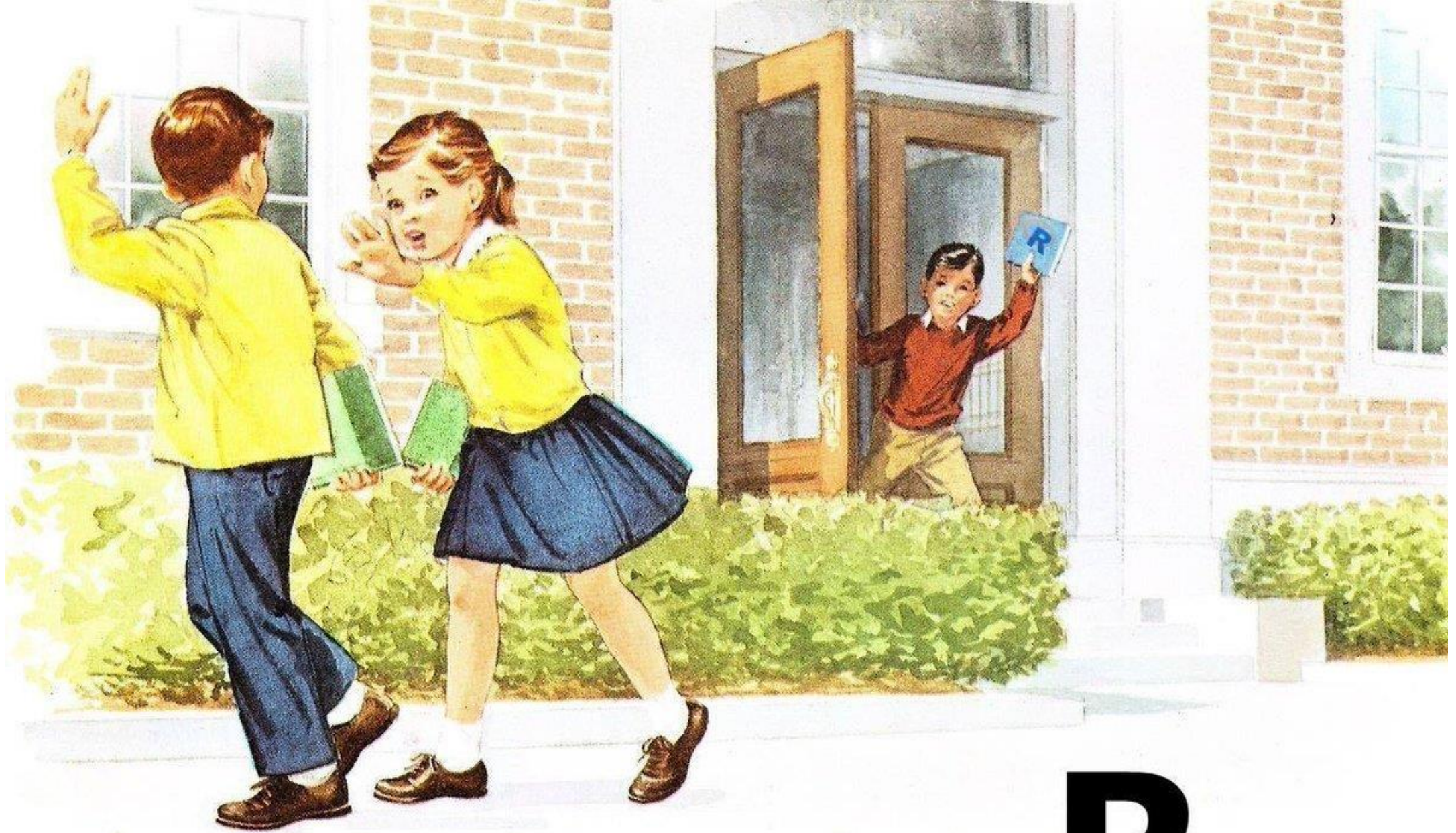
A: Define a project structure

B: Set a naming convention

C: Use scripted workflows

D: Digital notebooks

Reproducible
research



Run, or he's going to tell us about
again!

R

Still haven't found what I'm looking for

- Help your future-self

Final Final version








File Home Share View

← → ↕ ⬆ > This PC > B_Palmer_Medicine_Files > 4a Project > Pyrosequencing_analysis > Pyrosequencing_Paper > Draft_Paper_incl_Figs > Submission > JVI_Resubmission > JVI_resubmission_files > Final Final version

	Name	Date modified	Type	Size
✓ Quick access	Desktop			
	Downloads			
	Documents			
	Pictures			
	Projects			
	Google Drive			
	House			
	Google Drive File Stream (G:)			
	FAIR_workshop			
	Icon Files			
	R_Users_Workshop			
	subgroup_4-2_drafts			
	> OneDrive			
	✓ This PC			
		Cover_letter_B_A_Palmer_Sept_2014	10/09/2014 17:05	Microsoft Word 97 - 200...
	Fig_1_Sept_14	11/09/2014 10:31	Adobe Acrobat Docum...	25 KB
	Fig_1_Sept_14	10/09/2014 23:07	Microsoft PowerPoint 9...	158 KB
	Fig_2_Sept_14	11/09/2014 10:31	Adobe Acrobat Docum...	12 KB
	Fig_2_Sept_14	10/09/2014 23:07	Microsoft PowerPoint 9...	212 KB
	Fig_3_Sept_14	11/09/2014 10:31	Adobe Acrobat Docum...	173 KB
	Fig_3_Sept_14	10/09/2014 23:07	Microsoft PowerPoint 9...	527 KB
	Fig_4_Sept_14	11/09/2014 10:31	Adobe Acrobat Docum...	40 KB
	Fig_4_Sept_14	10/09/2014 23:07	Microsoft PowerPoint 9...	342 KB
	Fig_5_Sept_14	11/09/2014 10:33	Adobe Acrobat Docum...	12 KB
	Fig_5_Sept_14	10/09/2014 23:07	Microsoft PowerPoint 9...	178 KB
	HCV_UDPS_B_A_Palmer_Sept_14	17/09/2014 12:21	Microsoft Word 97 - 200...	442 KB
	Response_to_Reviewer_Sept_14	10/09/2014 22:42	Microsoft Word Docum...	559 KB
	Supplementary_Figure_B_A_Palmer_Sept_14	29/08/2014 13:21	Microsoft Word Docum...	378 KB
	Supplementary_Figure_B_A_Palmer_Sept_14	10/09/2014 22:31	Adobe Acrobat Docum...	224 KB
	Tables_B_A_Palmer_Sept_2014	10/09/2014 22:09	Microsoft Word 97 - 200...	185 KB

A: Define a generic project structure

This PC > Documents > Projects > generic.project > analysis

Name	Type
 data	File folder
 docs	File folder
 plots	File folder
 scripts	File folder
 tables	File folder
 generic	RMD File
 genericProject	R Project

B: Give your files and folders informative names

This PC > Documents > Projects > 2016-08-08_RespPCT > analysis > data

Name

Date modified



raw_data

21/01/2019 21:06



2018-11-06_abx

06/11/2018 13:10



2018-11-06_monitoring

06/11/2018 13:09



2018-11-06_pct

06/11/2018 13:08



2018-11-06_pt_info

06/11/2018 13:07

Outline a file naming convention

Machine readable:

- Inherent order
- Avoid spaces
- Avoid punctuation
- Remove case-sensitivity

Human readable:

- Contains info on content
- Avoid spaces
- Avoid punctuation
- Remove case sensitivity

Metadata:

Separate with underscores ("_")

- Avoid punctuation
- Remove case-sensitivity

01_`marshal-data`.r

02_`pre-dea-filtering`.r

03_`dea-with-limma-voom`.r

04_`explore-dea-results`.r

90_`limma-model-term-name-fiasco`.r

helper01_`load-counts`.r

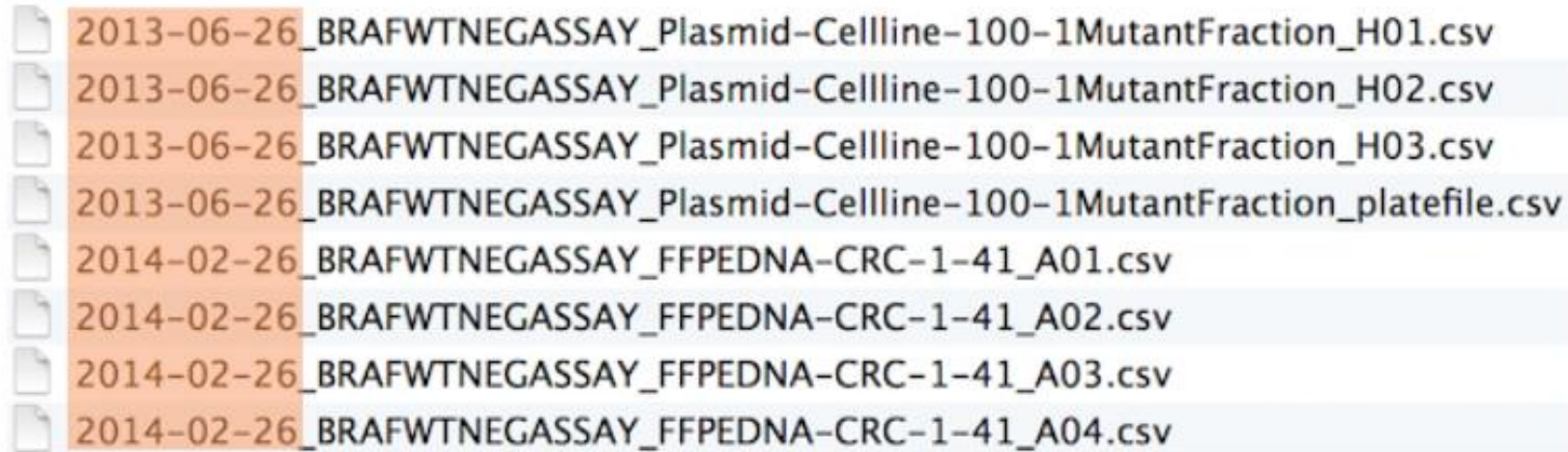
helper02_`load-exp-des`.r

helper03_`load-focus-statinf`.r

helper04_`extract-and-tidy`.r

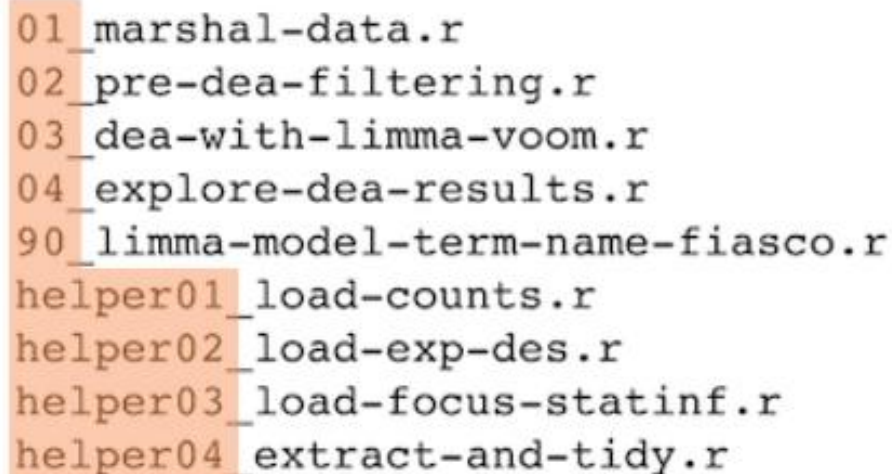
Outline a file naming convention

Chronological order:



```
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H01.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H02.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H03.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_platefile.csv
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A01.csv
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A02.csv
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A03.csv
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A04.csv
```

Logical order:














```
01_marshall-data.r
02_pre-dea-filtering.r
03_dea-with-limma-voom.r
04_explore-dea-results.r
90_limma-model-term-name-fiasco.r
helper01_load-counts.r
helper02_load-exp-des.r
helper03_load-focus-statinf.r
helper04_extract-and-tidy.r
```

Everything in its right place

- Make your file names:
 1. Machine readable
 2. Human readable
 3. Work with default ordering

NO









Name

-  All unique 4a amino acid Sequences (B-N).fas
-  All unique 4a amino acid Sequences (B-N).meg
-  All_AA_haplotypes.meg
-  All_AA_haplotypes_with_clonal_sequences.meg
-  BS100_AA_with_clones
-  BS100_AA_with_clones.nwk
-  BS1000_AA_pyro&clones
-  BS1000_AA_pyro&clones.nwk
-  BS1000_AA_pyro_only
-  BS1000_AA_pyro_only.nwk
-  BS1000_Unique_Clonal_AA

Yes

Projects > 2016-08-08_RespPCT > analysis > scripts

Name

-  01_clean_data
-  02_plots
-  03_tables
-  04_stats_analysis
-  05_post_hoc_stats
-  functions
-  randomization
-  tables

C: Joined up thinking

- The R scripts should also be human readable
 - Annotate the code
 - Break up the scripts into dedicated tasks
 - Interlink to other project scripts

```
1 # Data ----
2 # Eight tibbles returned from the 01_data_import_and_tidying_master_file.R
3 # 1. fgf23_data => FGF23 readings from study centres 01-03
4 # 2. food_level_data => Food diary entries
5 # 3. grouped_data => Dialysis and nondialysis diary entries by component
6 # 4. k_data => Serum potassium
7 # 5. master_data_clean => all the clean master file data if required
8 # 6. p_data => Serum phosphate
9 # 7. pth_data => Parathyroid hormone readings
10 # 8. pulses_nuts_data
11
12 source("scripts/01_data_import_and_tidying_master_file.R")
```


Work from the raw data ALWAYS!!



Tom Webb @tomjwebb · 16 Jan 2015

If you could tell a new PhD student one thing to help make their data more useful/shareable, what would it be?



27



11



7



Dr Gavin Simpson

@ucfagls

Follow

Replying to @tomjwebb

@tomjwebb don't, not even with a barge pole, not for one second, touch or otherwise edit the raw data files. Do any manipulations in script

7:15 AM - 16 Jan 2015

D: R Markdown

- R Markdown combines the code you wrote, the output produced and you own comments
- You can view it as a digital lab notebook, where you are both recording what you're doing, and what you were thinking while you were doing it!
- R Markdown outputs can take many forms
 - Word documents, PDFs, slideshows etc.
- Once created the .Rmd file get sent to knitr, which executes the chunks of code and creates a new markdown document
 - this is then processed by pandoc which creates the finished file
 - knitr and pandoc are external websites

R Markdown

YAML header

```
---  
title: "This is a reproducible document"  
date: 19th June 2019  
output: html_document  
---
```

Chunks of code

```
```{r setup, include = FALSE}  
library(ggplot2)
library(dplyr)
smaller <- diamonds %>%
filter(carat <= 2.5)
```
```

Plain text with data
outputs from R code

```
We have data about `r nrow(diamonds)`  
diamonds. Only  
`r nrow(diamonds) - nrow(smaller)` are  
larger than  
2.5 carats. The distribution of the  
remainder is shown below:
```

Chunks of code

```
```{r, echo = FALSE}  
smaller %>%
ggplot(aes(carat)) +
geom_freqpoly(binwidth = 0.01)
```
```


Install the Chrome plugin PubPeer

 NCBI Resources ▾ How To ▾ [Sign in to NCBI](#)

 PubMed ▾ [Create RSS](#) [Create alert](#) [Advanced](#) [Help](#)

Format: Abstract ▾ [Send to ▾](#)

See 1 citation found by title matching your search:

[J Surg Res.](#) 2019 Apr 26;241:235-239. doi: 10.1016/j.jss.2019.03.062. [Epub ahead of print]

 23 comments on PubPeer (by: Andrew D. Althouse, Thom Baguley, Guillaume A. Rousselet, Timothy Feeney, Paul M Brown, Frank E. Harrell, David Nunan, Samantha R. Seals, Raj Mehta, Yevgeniy Feyman, Ionomidotis Irregularis, Andrew Gelman, Aleks Reito, Daniel E. Leisman, Pavlos Msaouel, Ryan Miller, Maarten Van Smeden, Zad Rafi Chow)

Is the Power Threshold of 0.8 Applicable to Surgical Science?-Empowering the Underpowered Study.

[Bababekov YJ](#)¹, [Hung YC](#)², [Hsu YT](#)², [Udelsman BV](#)², [Mueller JL](#)², [Lin HY](#)², [Stapleton SM](#)², [Chang DC](#)².

 [Author information](#)

Abstract

BACKGROUND: Many articles in the surgical literature were faulted for committing type 2 error, or concluding no difference when the study was "underpowered". However, it is unknown if the current power standard of 0.8 is reasonable in surgical science.

Full text links

 FULL-TEXT ARTICLE

Save items



Similar articles



[Review](#) Interventions to Prevent Falls in Community- [Agency for Healthcare Research...]

[Review](#) Is There Truly "No Significant Difference"? Under [J Bone Joint Surg Am. 2015]

[Review](#) Randomized controlled trials and neurosurgery: the ideal fit or : [J Neurosurg. 2016]

[Review](#) Low-Dose Aspirin for the Prevention of Morbidity and [Agency for Healthcare Research...]

Further reading



Sam Westwood

@westwoodsam1

Following



I am embarking on my own [#PaperPerDayChallenge](#) where I read at least one paper, well, per day for a whole year. To kick start, [nature.com/articles/43573...](https://www.nature.com/articles/43573...) inspired by [@ukrepro](#) Reproducibility Workshop [@CumberlandLodge](#) and a talk by [@MarcusMunafo](#)



Scientists behaving badly

In a questionnaire-based survey of US biomedical researchers, respondents admitted to a range of dubious practices. Transgressions included failing to present data

[nature.com](#)

11:27 PM - 6 Jan 2019

Further reading



Sam Westwood

@westwoodsam1

Following

#119 I know I'm late to the party on this but Jesus! Pre-registration just tears up (inflated) effect sizes like a boss.

[frontiersin.org/articles/10.33 ...](https://frontiersin.org/articles/10.33)

frontiers
in Psychology

Specialty Section
published: 11 June 2019
doi: 10.3389/fpsyg.2019.00119

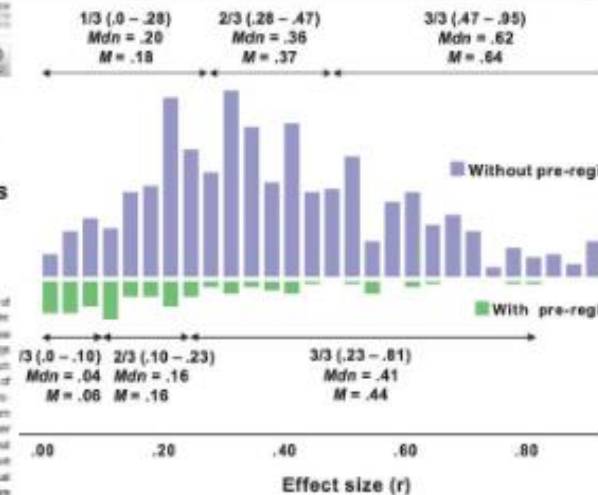
The Meaningfulness of Effect Sizes in Psychological Research: Differences Between Sub-Disciplines and the Impact of Potential Biases

Thomas Schuster¹ and Marcus A. Schwan²

¹Department of Psychology, University of Bamberg, Germany

Effect sizes are the currency of psychological research. They quantify the results of a study to answer the research question and are used to estimate statistical power. The interpretation of effect sizes—when is an effect small, medium, or large?—has been guided by the recommendations Jacob Cohen gave in his pioneering writings starting in 1988. Cohen compares an effect with the effects found in past research or use certain conventional benchmarks. The present analysis shows that neither of these recommendations is currently applicable. From past publications without pre-registration, 930 effects were randomly drawn and compared with 50 effects from publications with pre-registration, revealing a large difference (Effects from the former (median $r = 0.38$) were much larger than effects from the latter (median $r = 0.16$). That is, certain biases, such as publication bias or questionable research practices, have caused a dramatic inflation in published effects, making it difficult to compare an actual effect with the real population effects (as these are unknown). In addition, there were very large differences in the mean effects between psychological sub-disciplines and between different study designs, making it impossible to apply any global benchmarks. Many more pre-registered studies are needed in the future to derive a reliable picture of real population effects.

Keywords: effect size, Cohen's statistical power, sample size, replicability



(absolute values) from articles published with ($N = 82$) and without ($N = 684$) pre-registration transformed into a correlation coefficient r .

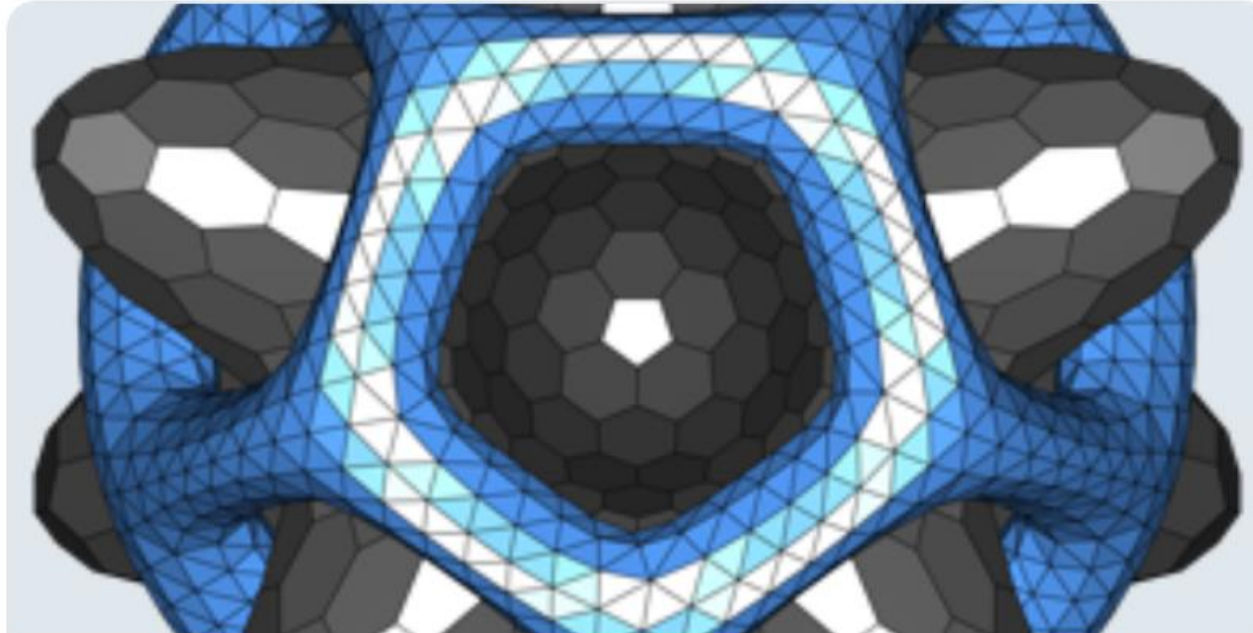
As an aside



Björn Brembs @brembs · Jun 17



Not tens of thousands of dollars - tens of dollars. This is what we should be paying for our scholarly articles: "Cost models for running an online open journal" buff.ly/2FeuSlv
[#oa](#) [#openaccess](#) [#OpenScience](#)



Cost models for running an online open journal | Journal of Open So...

Blog for the Journal of Open Source Software <http://joss.theoj.org>

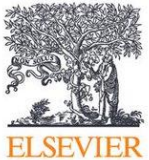
blog.joss.theoj.org

As an aside

Four principles of...



...the information system
supporting research



As an aside



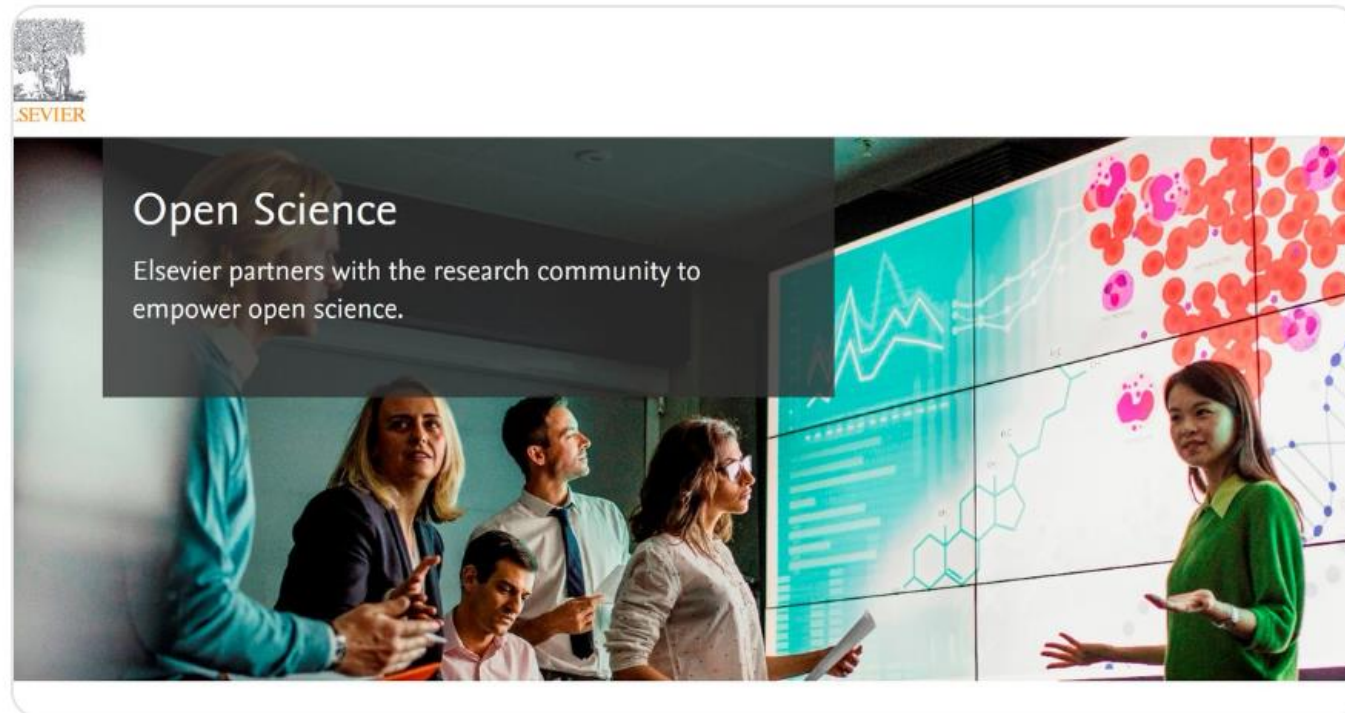
Jon Tennant ✓

@Protohedgehog

Following

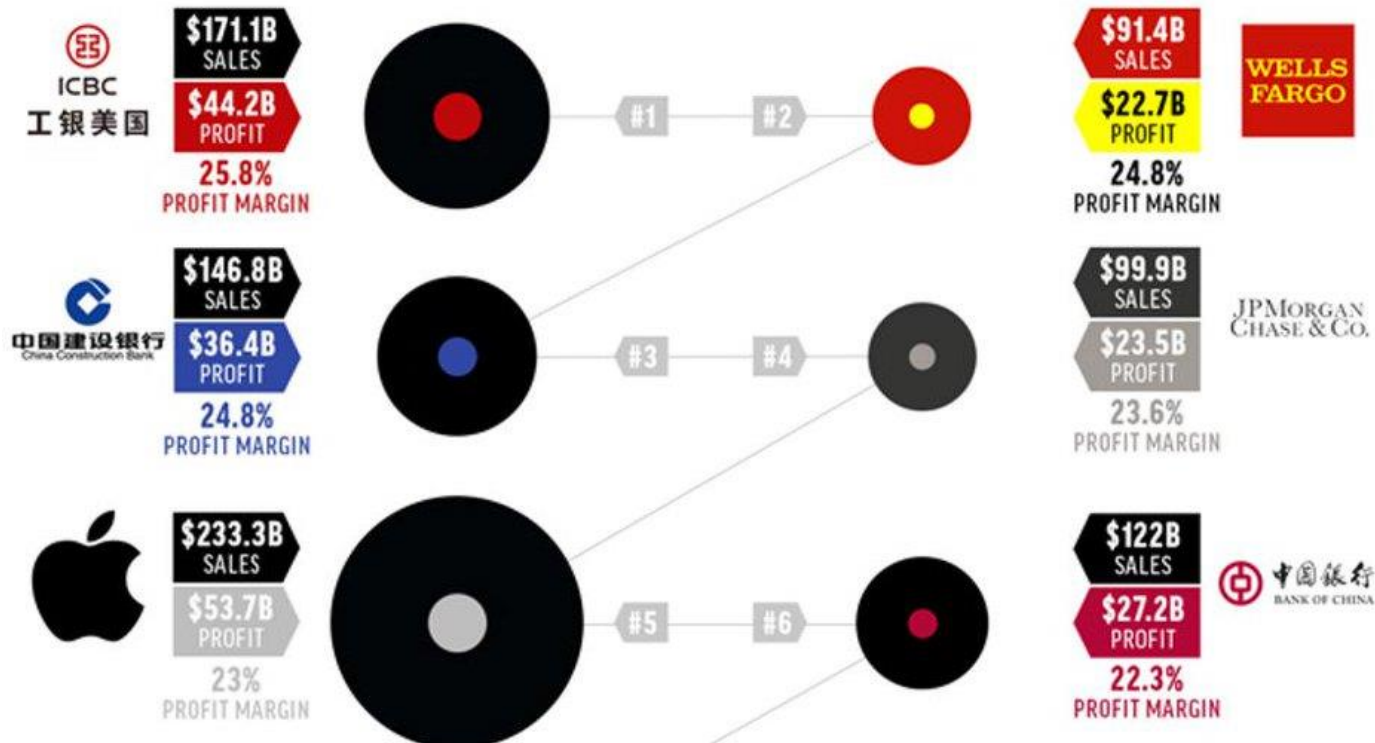


Elsevier saying that they "empower open science" is like McDonald's saying they empower vegetarians because they put gherkins in their cheeseburgers.



As an aside

FORBES GLOBAL 2000: PROFIT MARGINS



- For reference, Elsevier regularly make **> 30% profit**



Robert Thibault

@rt_thibault

Follow

Preprint on the cost of gold open access:

psyarxiv.com/hk4gn/

Last year's revenues from APCs: NPG > \$54m, Elsevier \$53m, Wiley \$42m, PLoS \$37m, Frontiers \$38m. Wiley's profit margin: 74%



The Rent's too High: Self-Archive for Fair Online Publicatio...

The main contributors of scientific knowledge—researchers—generally aim to disseminate their findings far and wide. And yet, publishing companies have largely kept these findings behind a...
psyarxiv.com

Twitter



UK Reproducibility Network

@ukrepro

UK Reproducibility Network: a peer-led consortium to investigate factors which contribute to robust research, provide training, and disseminate best practice.



Retraction Watch

@RetractionWatch

Tracking retractions as a window into the scientific process. Sign up for our daily newsletter: eepurl.com/bNRIUn Tips? team@retractionwatch.com



James Heathers

@jamesheathers

Research scientist. Biosignals, meta-science, error detection. Yelling has vitamins. Cohost of [@hertzpodcast](https://twitter.com/hertzpodcast).

📍 The warm embrace of Uncle Sam

🌐 jamesheathers.com



Brian Nosek

@BrianNosek

Executive Director @ Center for Open Science, Professor @ University of Virginia, and co-Founder of Project Implicit



Jon Tennant

@Protohedgehog

Rogue Palaeontologist currently at [@crisparis](https://twitter.com/crisparis). I do [@paleorXiv](https://twitter.com/paleorXiv), [@OpenScienceMOOC](https://twitter.com/OpenScienceMOOC), [@EGU_GC](https://twitter.com/EGU_GC) & [@PLOSspaleo](https://twitter.com/PLOSspaleo). Tweets snarky, irresponsible and my own. Do epic shit.

📍 Jurassic Park



Open Science MOOC

@OpenScienceMOOC Follows you

A community designed for students and researchers to help make 'Open' the default setting for the future of research. Slack: osmooc.herokuapp.com

📍 Everywhere



Malcolm Macleod #FBPE

@Maclomaclee Follows you

clinical neurologist, stroke trialist, and interested in improving the quality of laboratory research



Kate Button

@ButtonKate Follows you

Academic. Psychologist. Cognitive mechanisms of depression & anxiety. Meta-science & scientific rigour. Sporadic Twitterer.