

Correctly reporting your clinical trial

- Enhancing transparency and reproducibility

Brendan Palmer,

Clinical Research Facility - Cork &
School of Public Health, UCC





 Set status

Brendan Palmer

bapalmer

[Edit profile](#)

Twitter: @B_A_Palmer

🔗 <https://crfcsdau.github.io/about>

Organizations



Overview

Repositories 13

Projects 0

Stars 0

Followers 10

Following 8

Pinned

Customize your pins



1-day R workshop in using R-projects, writing cleaner code and a crash course in the tidyverse

● R ★



Short 1 hour introductions to R-related topics such as creating R projects, using GitHub through RStudio and more

● HTML ★ 1



Introductory R workshop: The tidyverse and beyond (6 x 2 hr sessions)

HTML 2 1



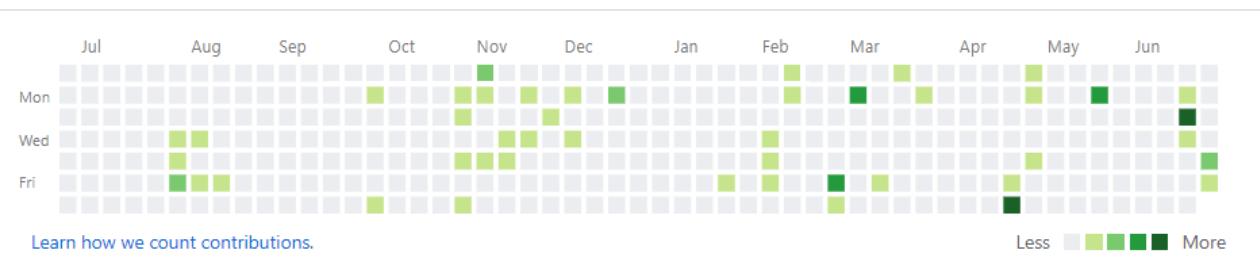
Section of the UCC Reproducible Conduct of Research digital badge dedicated to exposing researchers to reproducible research practices.

HTML

176 contributions in the last year

Contribution settings ▾

2019



COMMENTARY

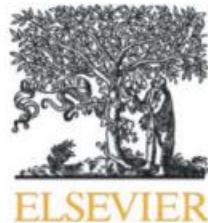
Scientists behaving badly

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

241 shades of grey



Contents lists available at SciVerse ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynim



Full Length Articles

The secret lives of experiments: Methods reporting in the fMRI literature

Joshua Carp

University of Michigan, Department of Psychology, 530 Church Street, Ann Arbor, MI, 48109, USA

ARTICLE INFO

Article history:
Accepted 3 July 2012
Available online 10 July 2012

Keywords:
fMRI
Methods reporting
Reproducibility
Experimental design
Analysis methods
Statistical power

ABSTRACT

Replication of research findings is critical to the progress of scientific understanding. Accordingly, most scientific journals require authors to report experimental procedures in sufficient detail for independent researchers to replicate their work. To what extent do research reports in the functional neuroimaging literature live up to this standard? The present study evaluated methods reporting and methodological choices across 241 recent fMRI articles. Many studies did not report critical methodological details with regard to experimental design, data acquisition, and analysis. Further, many studies were underpowered to detect any but the largest statistical effects. Finally, data collection and analysis methods were highly flexible across studies, with nearly as many unique analysis pipelines as there were studies in the sample. Because the rate of false positive results is thought to increase with the flexibility of experimental designs, the field of functional neuroimaging may be particularly vulnerable to false positives. In sum, the present study documented significant gaps in methods reporting among fMRI studies. Improved methodological descriptions in research reports would yield significant benefits for the field.

The key is in the detail

PLOS | BIOLOGY
FIFTEENTH ANNIVERSARY

BROWSE PUBLISH ABOUT

OPEN ACCESS

PERSPECTIVE

Risk of Bias in Reports of In Vivo Research: A Focus for Improvement

Malcolm R. Macleod , Aaron Lawson McLean, Aikaterini Kyriakopoulou, Stylianos Serghiou, Arno de Wilde, Nicki Sherratt, Theo Hirst, Rachel Hemblade, Zsanett Bahor, Cristina Nunes-Fonseca, Aparna Potluru, Andrew Thomson, Julija Baginskaitė, [...], Emily S. Sena [view all]

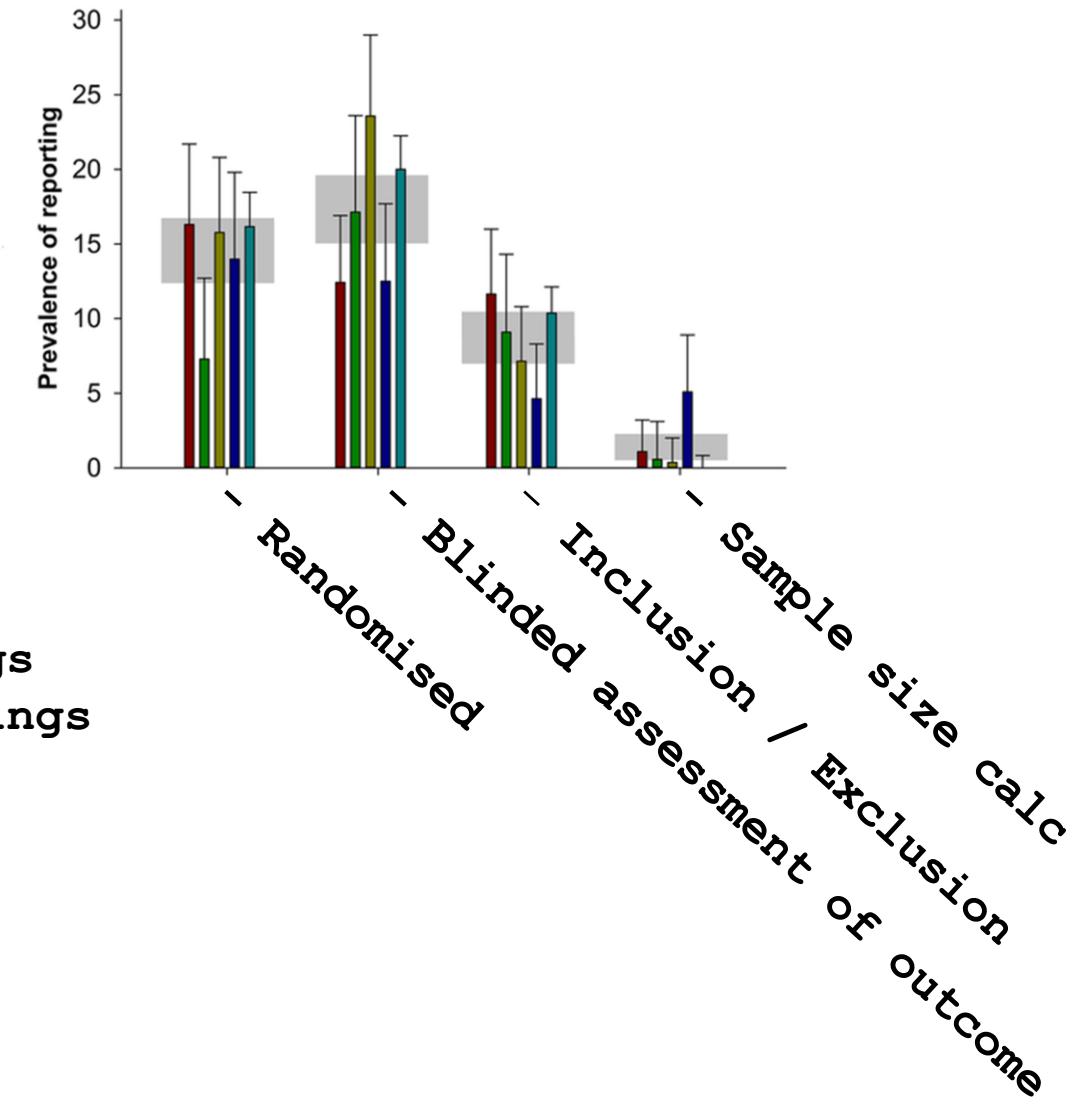
- 1,173 papers assessed
- Only one study did all four of these things
- 68% of studies did not do any of these things



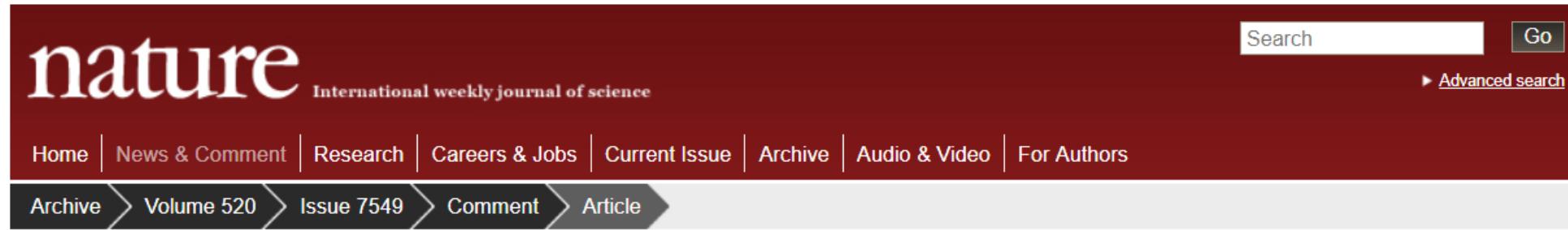
Original Article |  Full Access

1,026 Experimental treatments in acute stroke

Victoria E. O'Collins B.Sc., Malcolm R. Macleod MRCP, PhD, Geoffrey A. Donnan MD, FRACP, Laura L. Horky MD, PhD, Bart H. van der Worp MD, PhD, David W. Howells PhD 



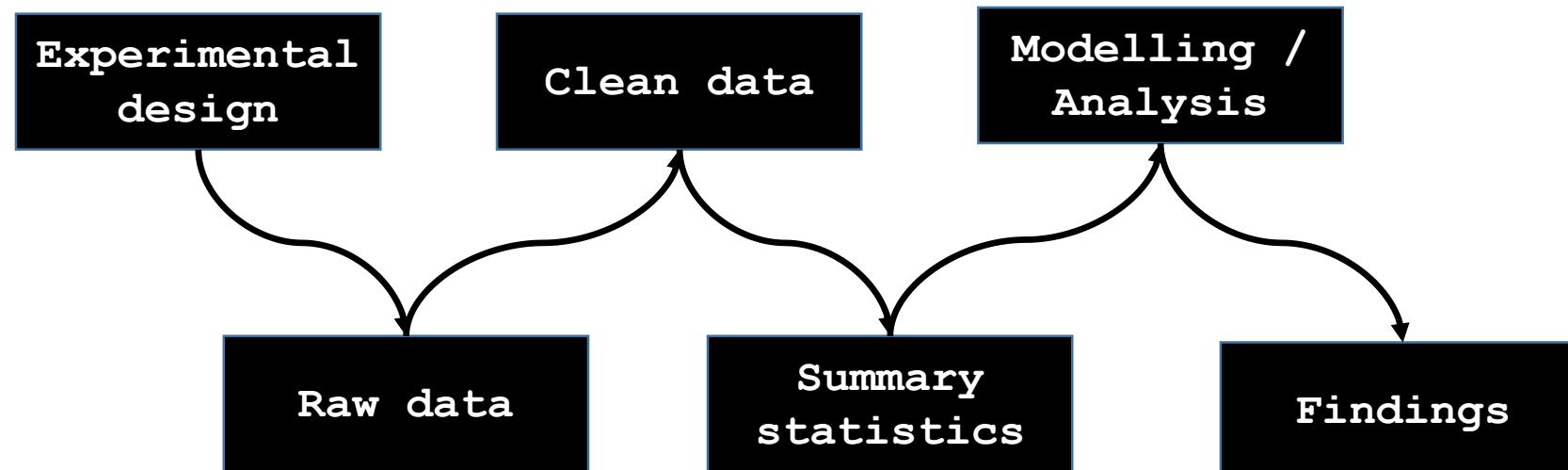
p-values should not define a study

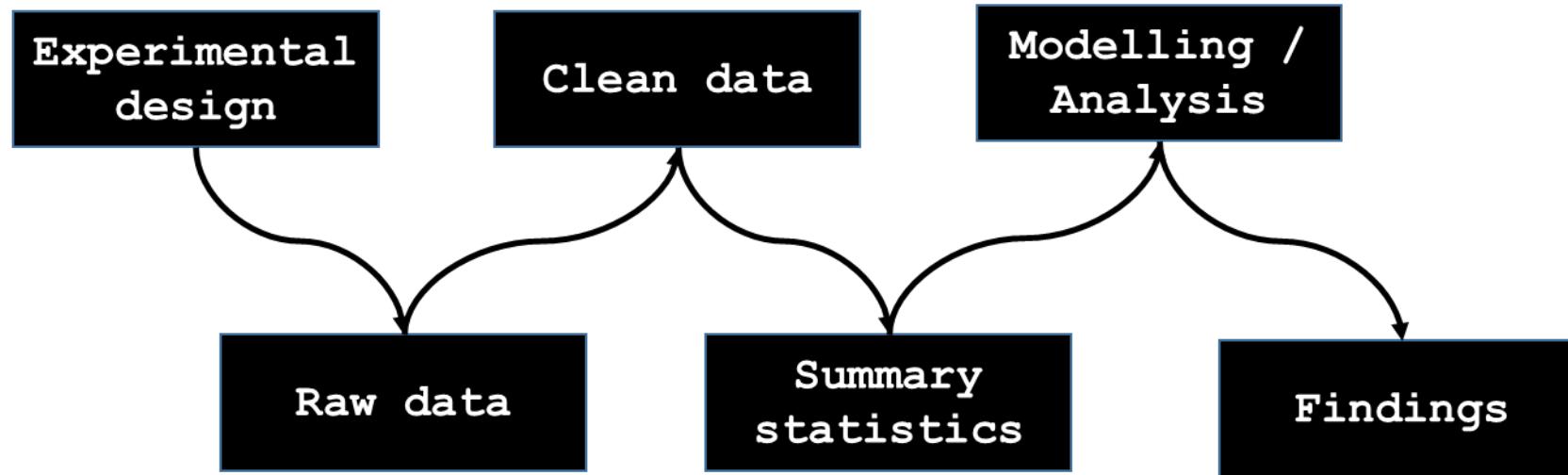


Statistics: *P* values are just the tip of the iceberg

Jeffrey T. Leek & Roger D. Peng

28 April 2015





Little debate

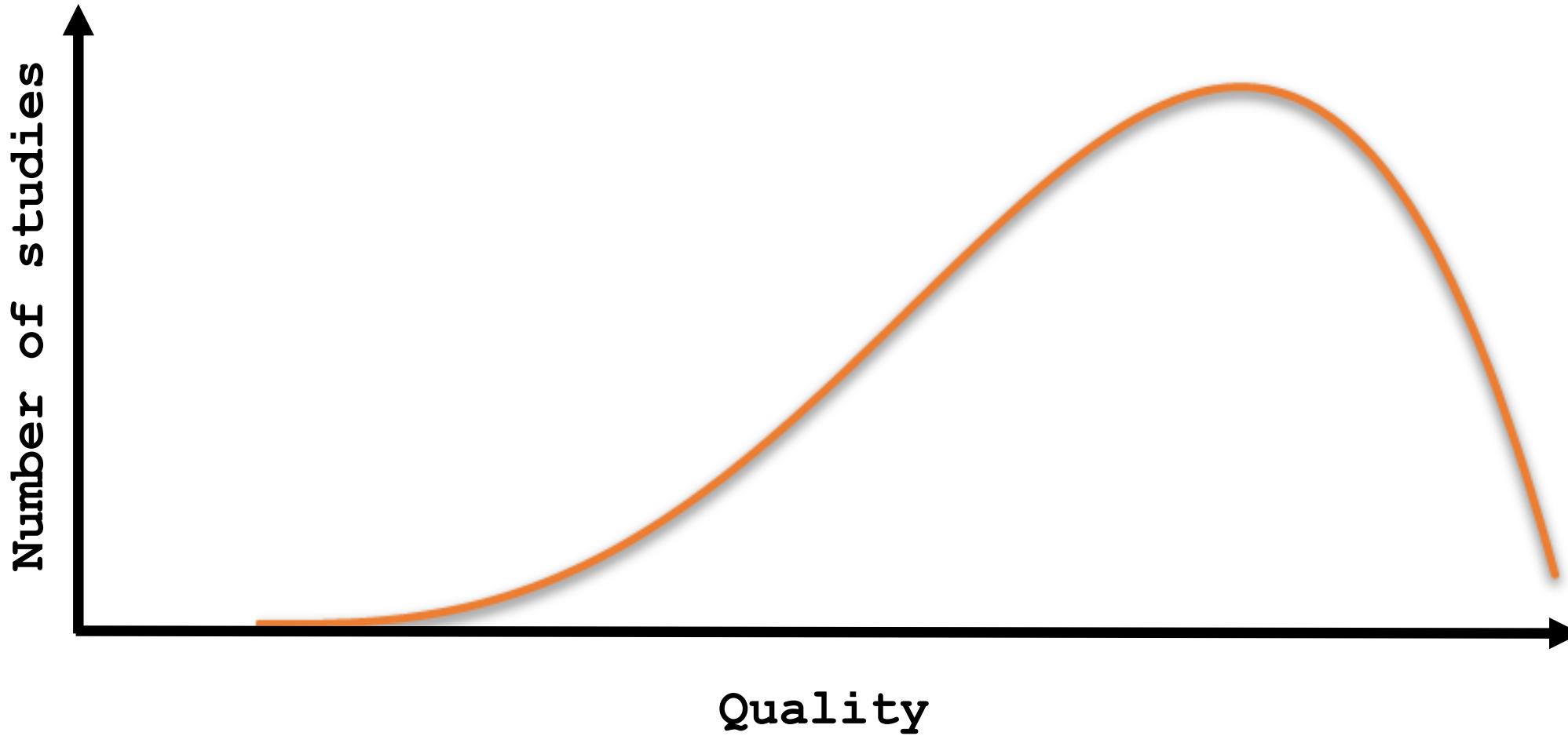
p-value

Extreme
scrutiny

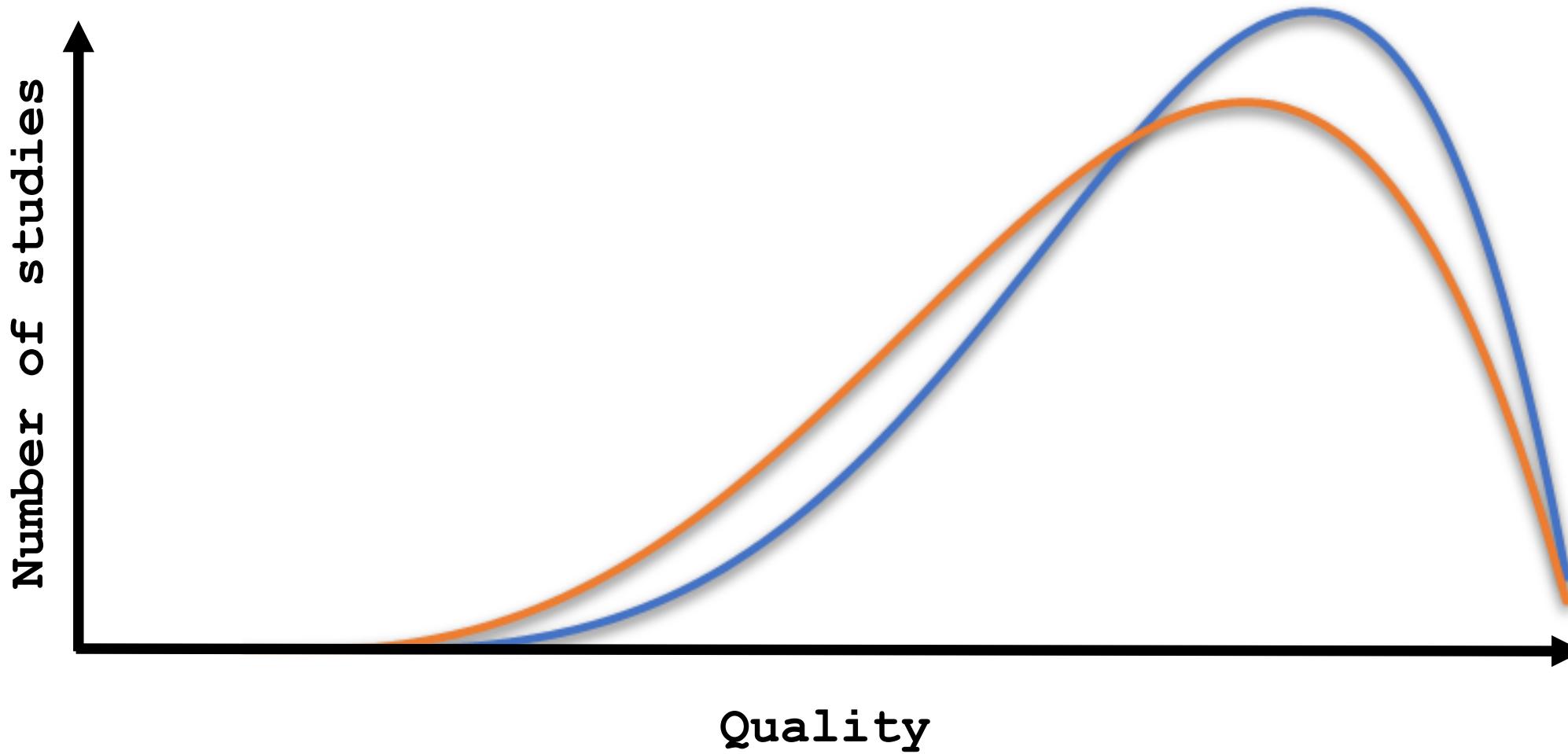


p-value

Today



Tomorrow



Who benefits most from reproducibility?



Casey Greene
@GreeneScientist

Follow

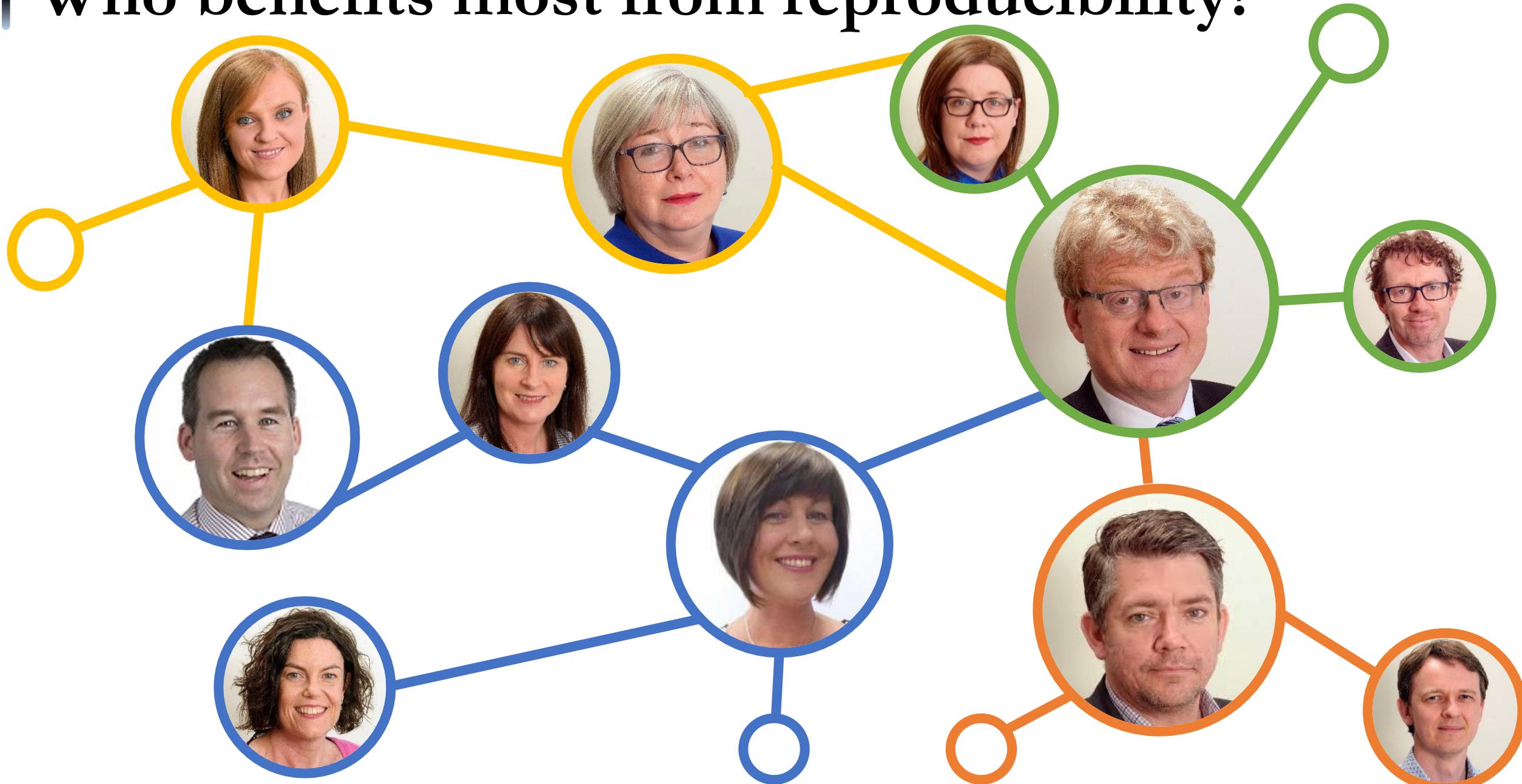


Reproducibility is important because the you
of 3 months ago is terrible at answering
email! - [@tracykteal](#) at [#2016dssummit](#)

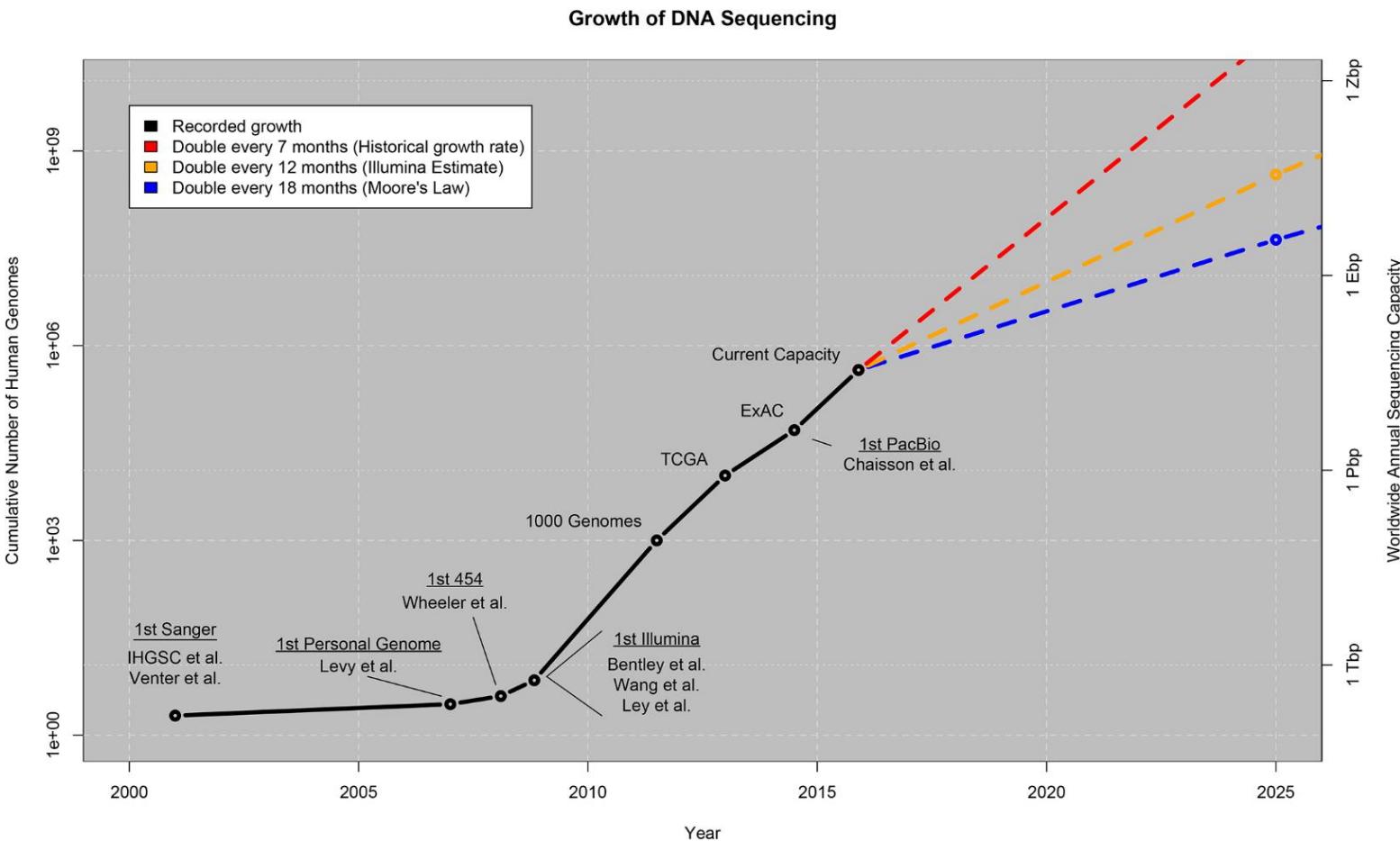
1:17 PM - 26 Oct 2016 from [Manhattan, NY](#)



Who benefits most from reproducibility?



The challenge



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



Where to begin...



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-

Live Home Page Apple Apple Support Apple Store iTools Mac OS X Microsoft MacTopics

NCBI LocusLink

PubMed Entrez BLAST OMIM Taxonomy Structure

Search LocusLink Display Brief Organism: All

Query: Go Clear

View Hs NEDD5 One of 1 Loci Save All Loci

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

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

PUB OMIM A VIEW UNIGENE MAP VAR HOMOL GDB

e! 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, hNed5, KIAA0158

Relationships ?

Map

RefSeq

GenBank

Links

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 ?

Less stress, more success

	A	B	C	D	E	F	G	H	I	J	K	L
1	id	week_no	filter_nam	treatment	replicate_no	flavonoids	biomass	variety	date	investigator		
2	1	0	ptp	nofilter	1	1.061	0.39	cos	2019/04/01	Darren Dahly		
3	2	0	ptp	nofilter	2	1.1805	0.42	cos	2019/04/01	Darren Dahly		
4	3	0	ptp	nofilter	3	1.0345	0.62	cos	2019/04/01	Darren Dahly		
5	4	0	ptp	nofilter	4	1.094	0.63	cos	2019/04/01	Brendan Palmer		
6	1	0	my	nofilter	1	1.061	0.39	cos	2019/04/01	Brendan Palmer		
7	2	0	my	nofilter	2	1.1805	0.42	cos	2019/04/01	Brendan Palmer		
8	3	0	my	nofilter	3	1.0345	0.62	cos	2019/04/01	Brendan Palmer		
9	4	0	my	nofilter	4	1.094	0.63	cos	2019/04/01	Brendan Palmer		
10	1	0	ca	nofilter	1	1.061	0.39	cos	2019/04/01	Brendan Palmer		
11	2	0	ca	nofilter	2	1.1805	0.42	cos	2019/04/01	Brendan Palmer		
12	3	0	ca	nofilter	3	1.0345	0.62	cos	2019/04/01	Brendan Palmer		
13	4	0	ca	nofilter	4	1.094	0.63	cos	2019/04/01	Darren Dahly		
14	5	1	ptp	filter	1	0.87	0.76	cos	2019/04/08	Darren Dahly		
15	6	1	ptp	filter	2	0.847	0.95	cos	2019/04/08	Darren Dahly		
16	7	1	ptp	filter	3	1.022	0.95	cos	2019/04/08	Darren Dahly		
17	8	1	ptp	filter	4	0.916	0.95	cos	2019/04/08	Darren Dahly		
18	9	1	my	filter	1	1.119	1.55	cos	2019/04/08	Darren Dahly		
19	10	1	my	filter	2	0.845	3.16	cos	2019/04/08	Darren Dahly		
20	11	1	my	filter	3	1.299	4.9	cos	2019/04/08	Brendan Palmer		
21	12	1	my	filter	4	1.149	5.5	cos	2019/04/08	Brendan Palmer		
22	13	1	ca	filter	1	0.716	5.5	cos	2019/04/08	Brendan Palmer		
23	14	1	ca	filter	2	0.881	7.94	cos	2019/04/08	Brendan Palmer		
24	15	1	ca	filter	3	0.586	8.71	cos	2019/04/08	Brendan Palmer		
25	16	1	ca	filter	4	0.561	8.71	cos	2019/04/08	Brendan Palmer		
26	17	2	ptp	filter	1	0	14.45	cos	2019/04/15	Brendan Palmer		
27	18	2	ptp	filter	2	1.006	2.14	cos	2019/04/15	Brendan Palmer		
28	19	2	ptp	filter	3	1.236	1.86	cos	2019/04/15	Brendan Palmer		
29	20	2	ptp	filter	4	1.206	1.2	cos	2019/04/15	Brendan Palmer		
30	21	2	mv	filter	1	1.545	2.45	cos	2019/04/15	Brendan Palmer		

data

dictionary

values



Less stress, more success

Less stress, more success

1	A	B	C	D	E	F	G	H	I	J	K	L
2	1	0	ptp	nofilter	1	1.061	0.39	cos	2019/04/01	Darren Dahly		
3	2	0	ptp	A	B	C	D	E				
4	3	0	ptp	1	field_name	data_type	data_format	example	standard_units	description		
5	4	0	ptp	2	id	numeric	integer	23	NA	Unique identifier applied to each observation		
6	1	0	my	3	week_no	numeric	integer					
7	2	0	my	4	filter_name	character	NA					
8	3	0	my	5	treatment	character	NA					
9	4	0	my	6	replicate_no	numeric	integer					
10	1	0	ca	7	flavonoids	numeric	double					
11	2	0	ca	8	biomass	numeric	double					
12	3	0	ca	9	variety	character	NA					
13	4	0	ca	10	date	date	YYYY/MM/DD					
14	5	1	ptp	11	investigator	character	Firstname Lastname					
15	6	1	ptp	12								
16	7	1	ptp	13								
17	8	1	ptp	14								
18	9	1	my	15								
19	10	1	my	16								
20	11	1	my	17								
21	12	1	my	18								
22	13	1	ca	19								
23	14	1	ca	20								
24	15	1	ca	21								
25	16	1	ca	22								
26	17	2	ptp	23								
27	18	2	ptp	24								
28	19	2	ptp	25								
29	20	2	ptp	26								
30	21	2	mv	27								
		data	dictionary	28								
				29								
				30								

The screenshot shows a data entry interface with two tabs: 'data' and 'dictionary'. The 'data' tab displays a grid of experimental data. The 'dictionary' tab provides a detailed schema for each column, including field_name, data_type, data_format, example, standard_units, and description. A tooltip for 'id' indicates it is a unique identifier applied to each observation.

Below the tabs, there are navigation buttons for the data grid: back, forward, data, dictionary, values, and a plus sign.

Less stress, more success

Step by step guide

← → C ⌂ 🔒 https://www.youtube.com/watch?v=Ry2xjTBtNFE

YouTube IE

Search

Book1 - Excel (Product Activation Failed)

Dahly, Darren Share

File Home Insert Page Layout Formulas Data Review View Tell me what you want to do...

Cut Copy Format Painter

Font Alignment Number Styles Cells Editing

D2

1	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
2	id	gender	gender_other	age	nationality	year_program																
3																						
4																						
5																						
6																						
7																						
8																						
9																						
10																						
11																						
12																						
13																						
14																						
15																						
16																						
17																						
18																						
19																						
20																						
21																						
22																						
23																						
24																						
25																						
26																						
27																						
28																						

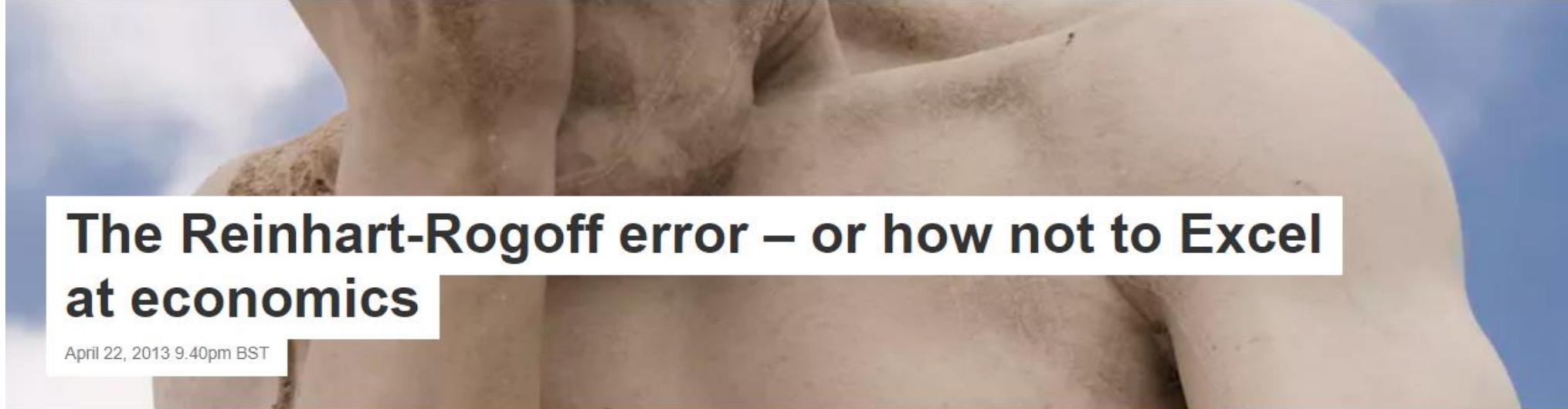
Sheet1 Sheet2 Sheet3 +

Ready

Type here to search



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](#)

88

[Facebook](#)

453

[LinkedIn](#)

[Print](#)

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.

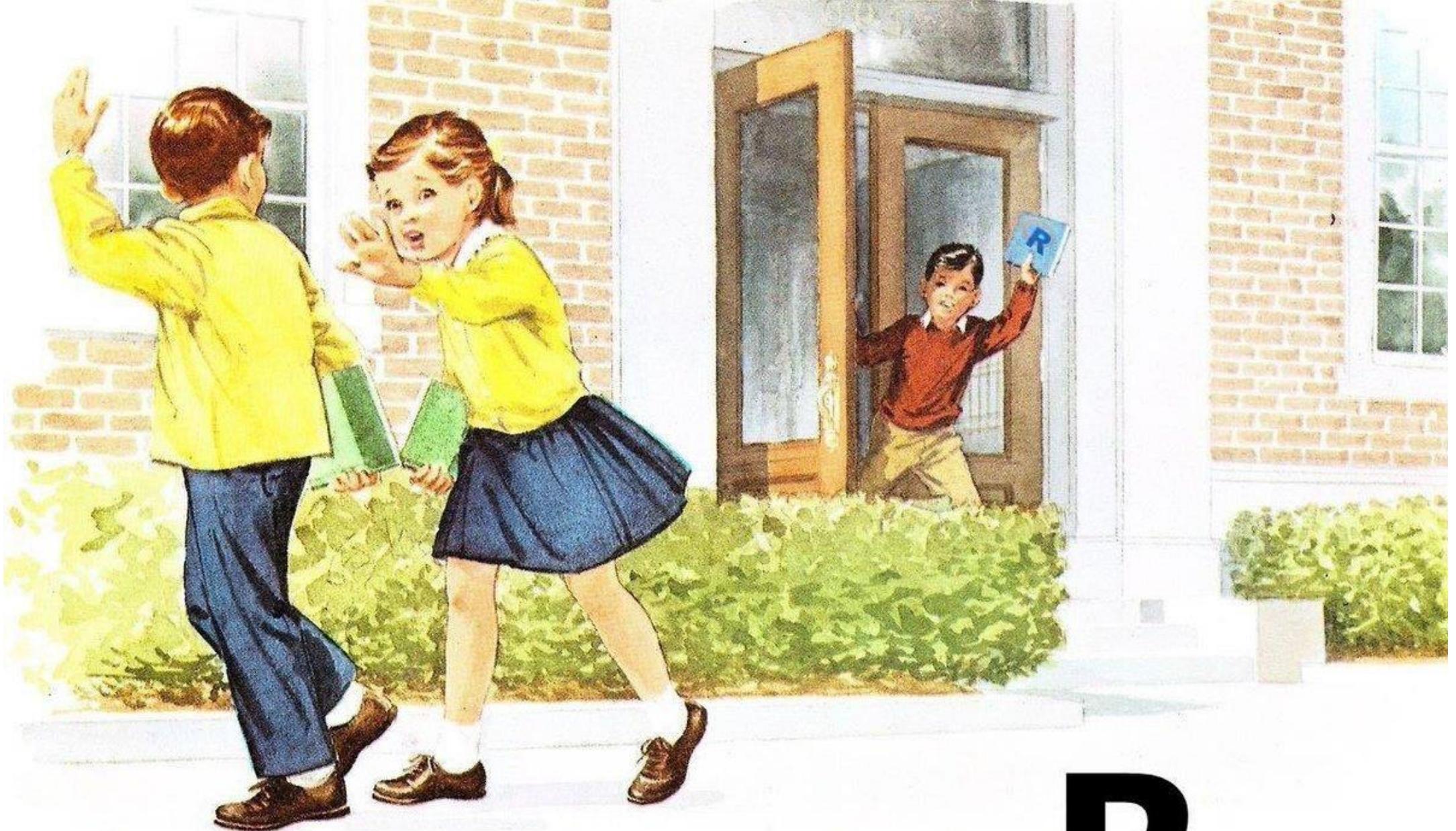
Sadly, these are not the first mistakes of this size and nature when handling data. So what on Earth went wrong, and can we fix it?

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

Putting the pieces together

- A: Define a project structure
- B: Set a naming convention
- C: Use scripted workflows
- D: Digital reports

Reproducible
research



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

R

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: 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_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

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

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_Uncle_Clonal_AA

Yes

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

Name
R 01_clean_data
R 02_plots
R 03_tables
R 04_stats_analysis
R 05_post_hoc_stats
R functions
R randomization
R 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 your 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 gets 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

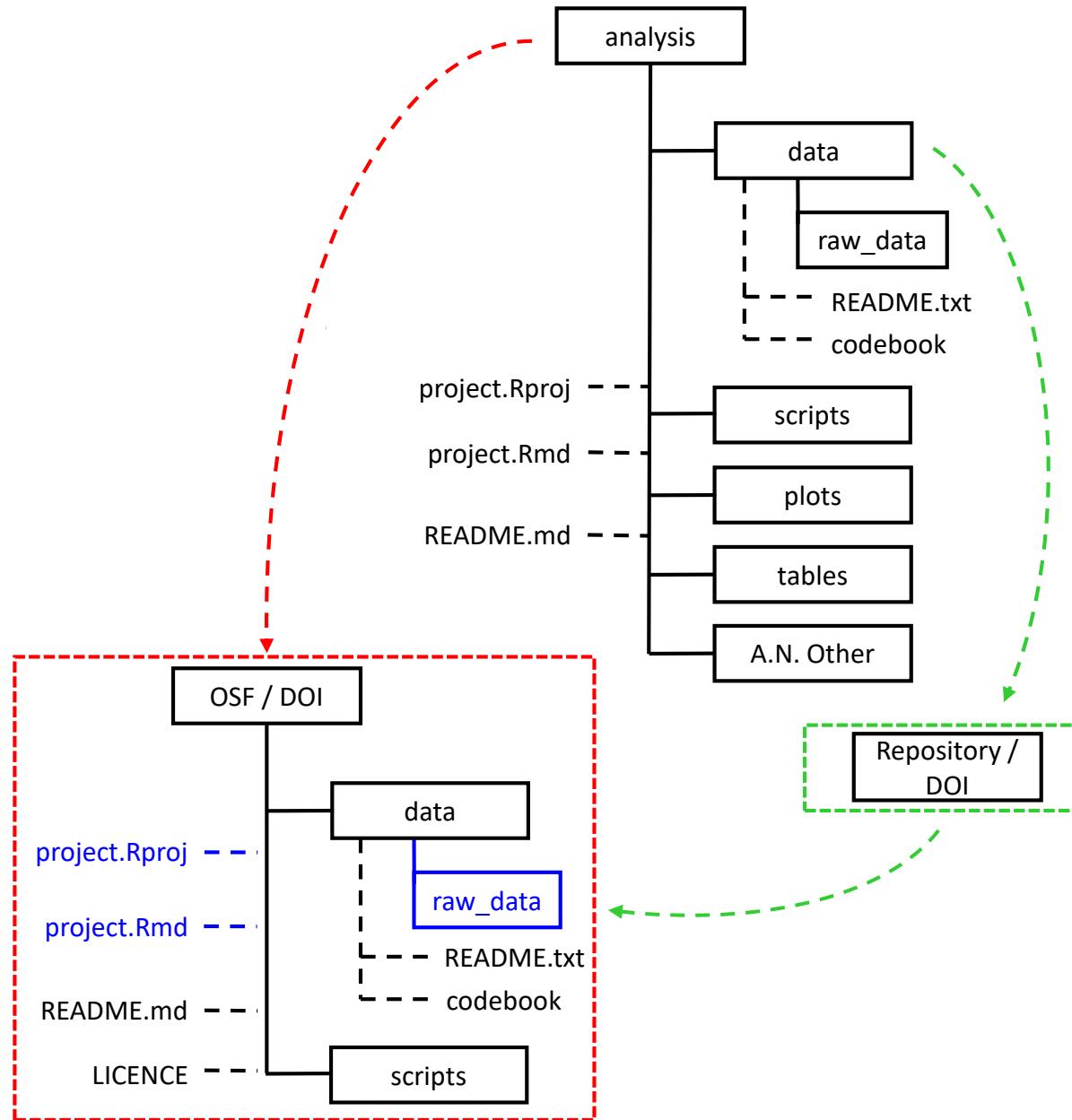
We have data about `r nrow(diamonds)` diamonds. Only `r nrow(diamonds) - nrow(smaller)` are larger than 2.5 carats. The distribution of the remainder is shown below:

Chunks of code

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

```
```
```

What does this allow us to do?



Try it for yourself

A screenshot of a web browser displaying the RStudio Cloud homepage. The URL in the address bar is <https://rstudio.cloud>. The page features a large blue header with the RStudio Cloud logo and navigation links for Log In and Sign Up. Below the header, a large white cloud shape contains the text "Welcome to RStudio Cloud alpha" and "Do, share, teach and learn data science with R.". A green "Get Started" button is positioned within the cloud. At the bottom of the page, a note states: "If you already have an RStudio shinyapps.io account, you can log in using your existing credentials."

R Studio Cloud

Log In Sign Up ☰

Welcome to RStudio Cloud ^{alpha}

Do, share, teach and learn data science with R.

Get Started

If you already have an RStudio shinyapps.io account, you can log in using your existing credentials.

Try it for yourself

https://rstudio.cloud/project/140507

Your Workspace / RCR digital badge

Brendan Palmer

File Edit Code View Plots Session Build Debug Profile Tools Help

Addins

03_final_analysis.R

```
36 labs(x = "",  
37   y = "Flavonoids (ppm)",  
38   title = "Lettuce variety vs flavonoid content") +  
39 theme(panel.background = element_blank(),      #Remove grey background  
40       # panel.grid.minor = element_blank(),  
41       axis.title = element_text(face = "bold", size = 12),  
42       axis.text = element_text(face = "bold", size = 10),  
43       axis.line = element_line(colour = "black", size = 1),  
44       plot.title = element_text(hjust = 0.0))  
45  
46 report_plot  
47  
48 # ggsave(paste("plots/", Sys.Date(), "_final_plot.png",  
49 #           sep = ""),  
50 #           report_plot,  
51 #           width = 16,  
49:1 # Read in the clean lettuce data
```

Environment History Connections Git

Import Dataset

Global Environment

Data

| | |
|-------------|--------------------------|
| data | 144 obs. of 10 variables |
| report_plot | List of 9 |
| test_data | 36 obs. of 10 variables |

Values

| | |
|-----------------|--|
| lettuce_variety | Named chr [1:3] "Cos Dixter" "Red Oakleaf" "Skyphos" |
|-----------------|--|

Files Plots Packages Help Viewer

Zoom Export

Lettuce variety vs flavonoid content

Flavonoids (ppm)

Cos Dixter Skyphos Red Oakleaf

Install the Chrome plugin PubPeer

NCBI Resources ▾ How To ▾ Sign in to NCBI

PubMed ▾ Is the Power Threshold of 0.8 Applicable to Surgical Science? Search

PubMed.gov US National Library of Medicine National Institutes of Health 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, Aleksi 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 YJ¹, Hung YC², Hsu YT², Udelsman BV², Mueller JL², Lin HY², Stapleton SM², Chang DC².

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 ELSEVIER FULL-TEXT ARTICLE

Save items

Similar articles

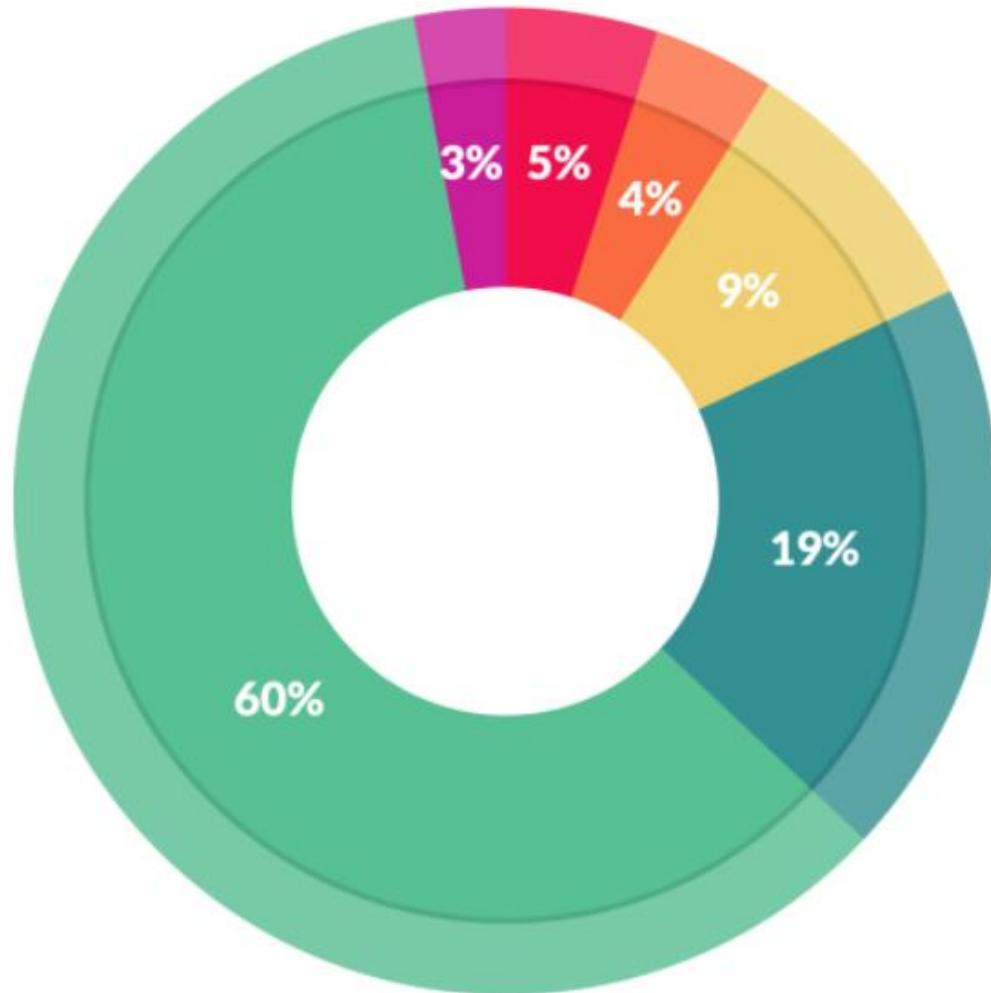
Review Interventions to Prevent Falls in Community-L Agency for Healthcare Research...]

Review Is There Truly "No Significant Difference"? Underl 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 anc Agency for Healthcare Research...]

It costs a lot of money to run a clinical trial



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

The best time to plant
a tree was 20 years ago

The second best time
is now

