Correctly reporting your clinical trial

- Enhancing transparency and reproducibility

Brendan Palmer,
Clinical Research Facility - Cork &
School of Public Health, UCC
@B_A_Palmer

CRF-C
HRB Clinical Research Facility Cork

Health Research Board
TMRN
Trials Methodology Research Network

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh
School of
Public Health

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/ynimg

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

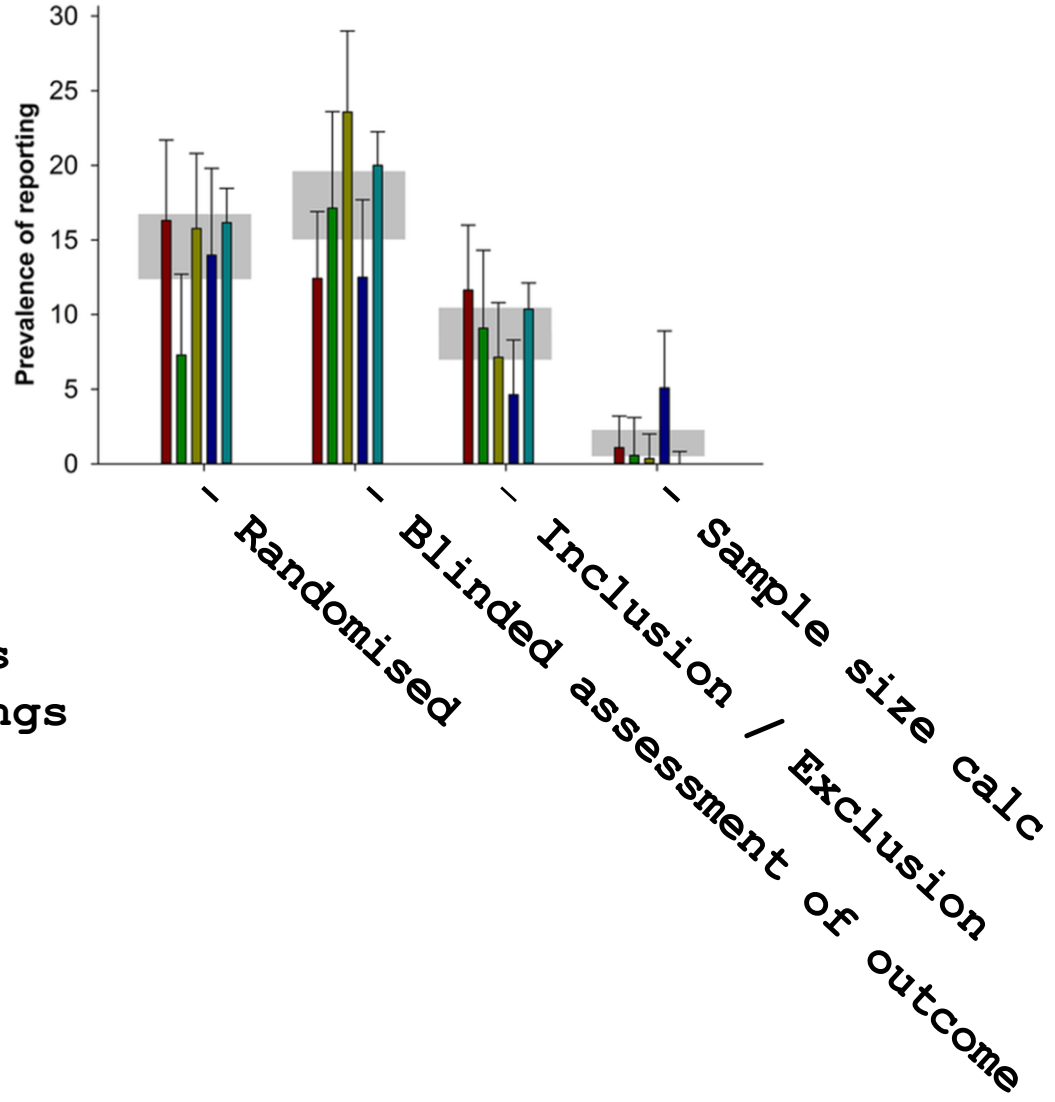
BROWSE PUBLISH ABOUT

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 Baginskitae, [...], 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

Annals of NEUROLOGY
An Official Journal of the American Neurological Association and the Child Neurology Society
ANA AMERICAN NEUROLOGICAL ASSOCIATION

Original Article | Full Access

1,026 Experimental treatments in acute stroke

Victoria E. O'Collins B.Sci, 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

nature International weekly journal of science

Search Go

Advanced search

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

Archive > Volume 520 > Issue 7549 > Comment > Article

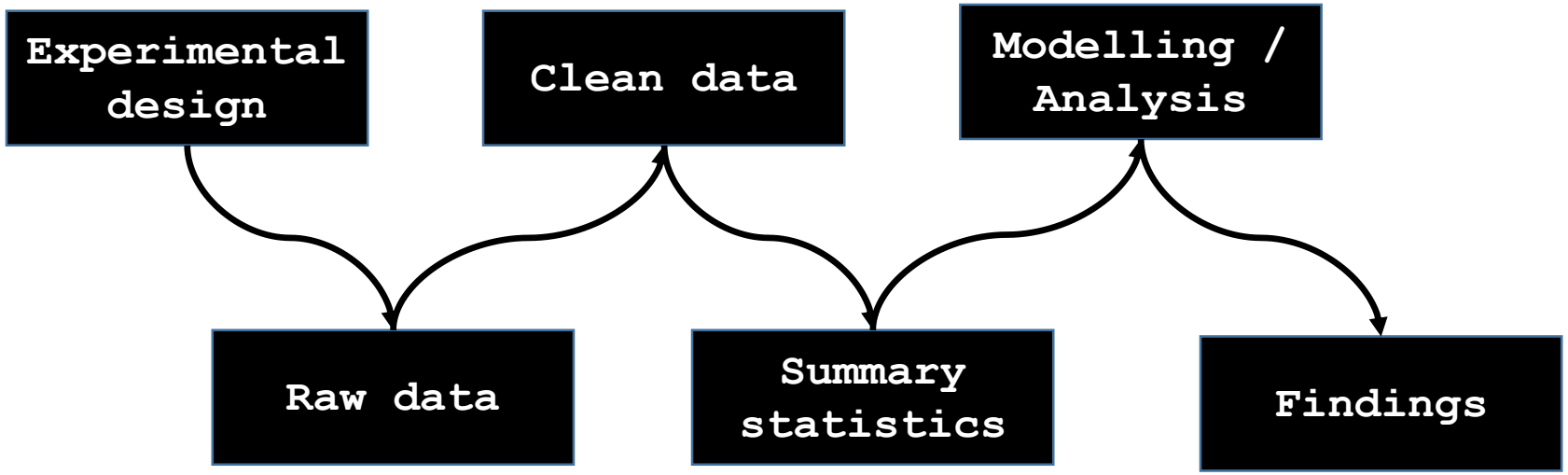
E-alert RSS Facebook Twitter

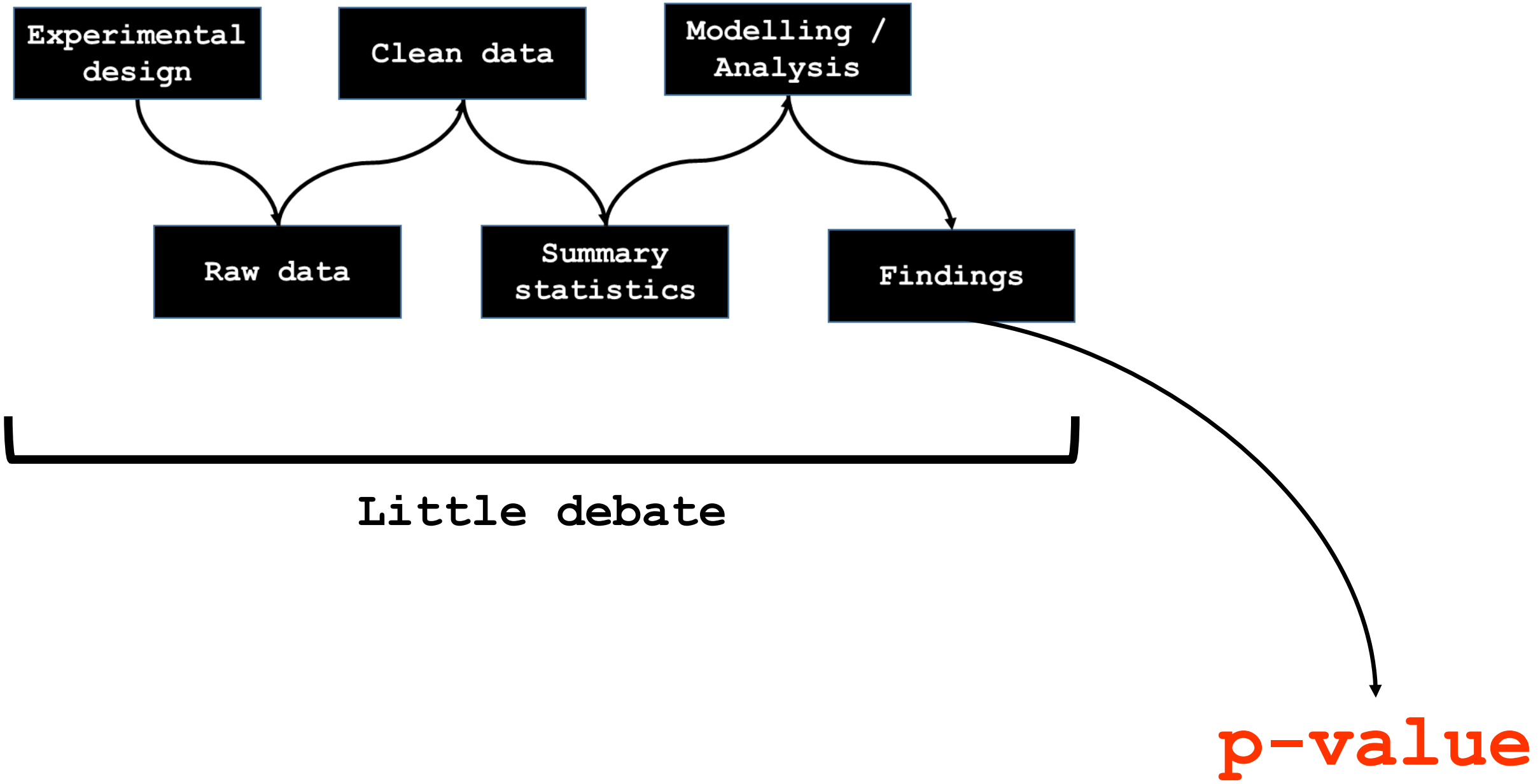
NATURE | COMMENT

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

Jeffrey T. Leek & Roger D. Peng

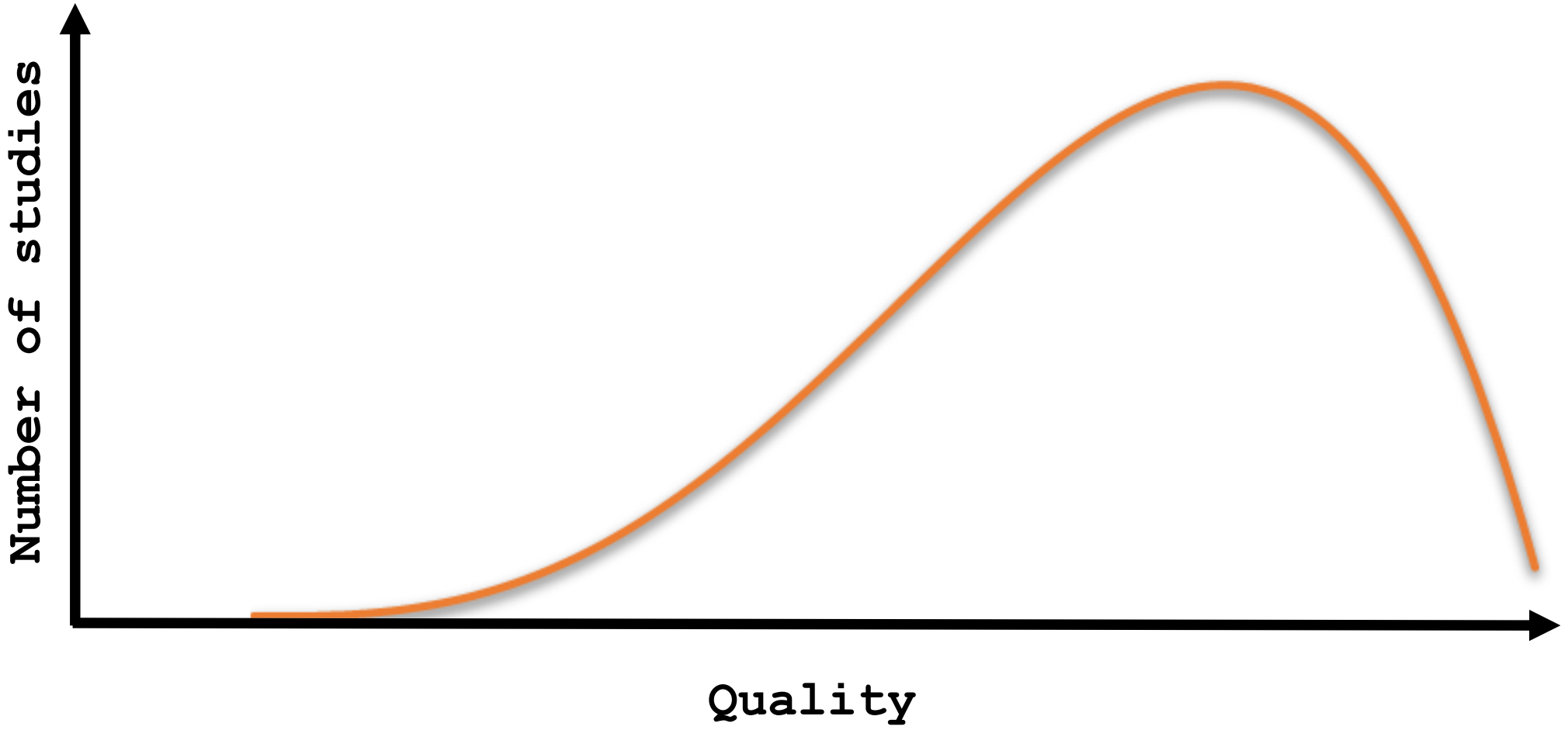
28 April 2015



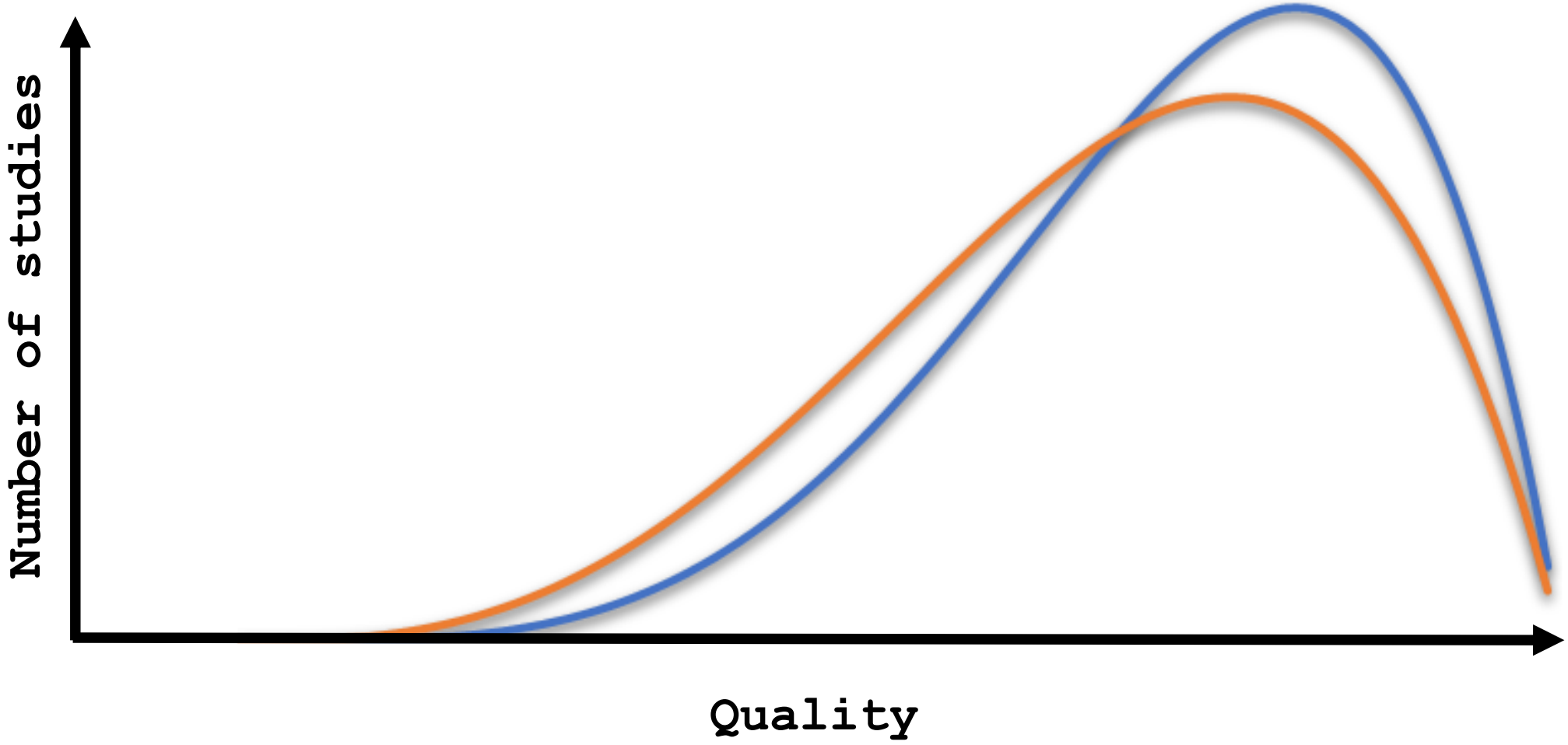

Experimental design
Clean data
Modelling / Analysis
Raw data
Summary statistics
Findings
Little debate
p-value

Extreme scrutiny

Today



Tomorrow



Who benefits most from reproducibility?

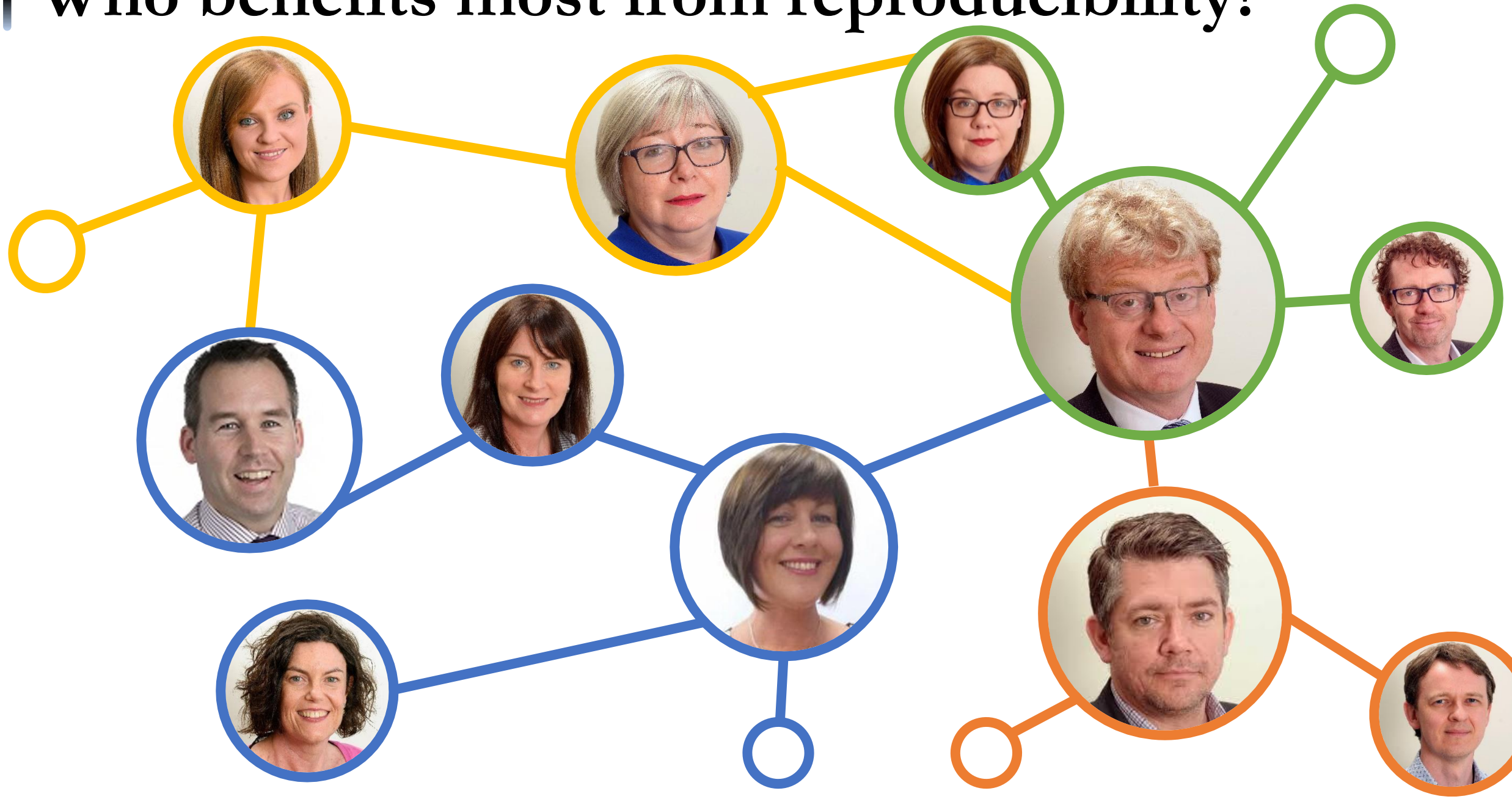
Casey Greene
@GreeneScientist

Follow

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

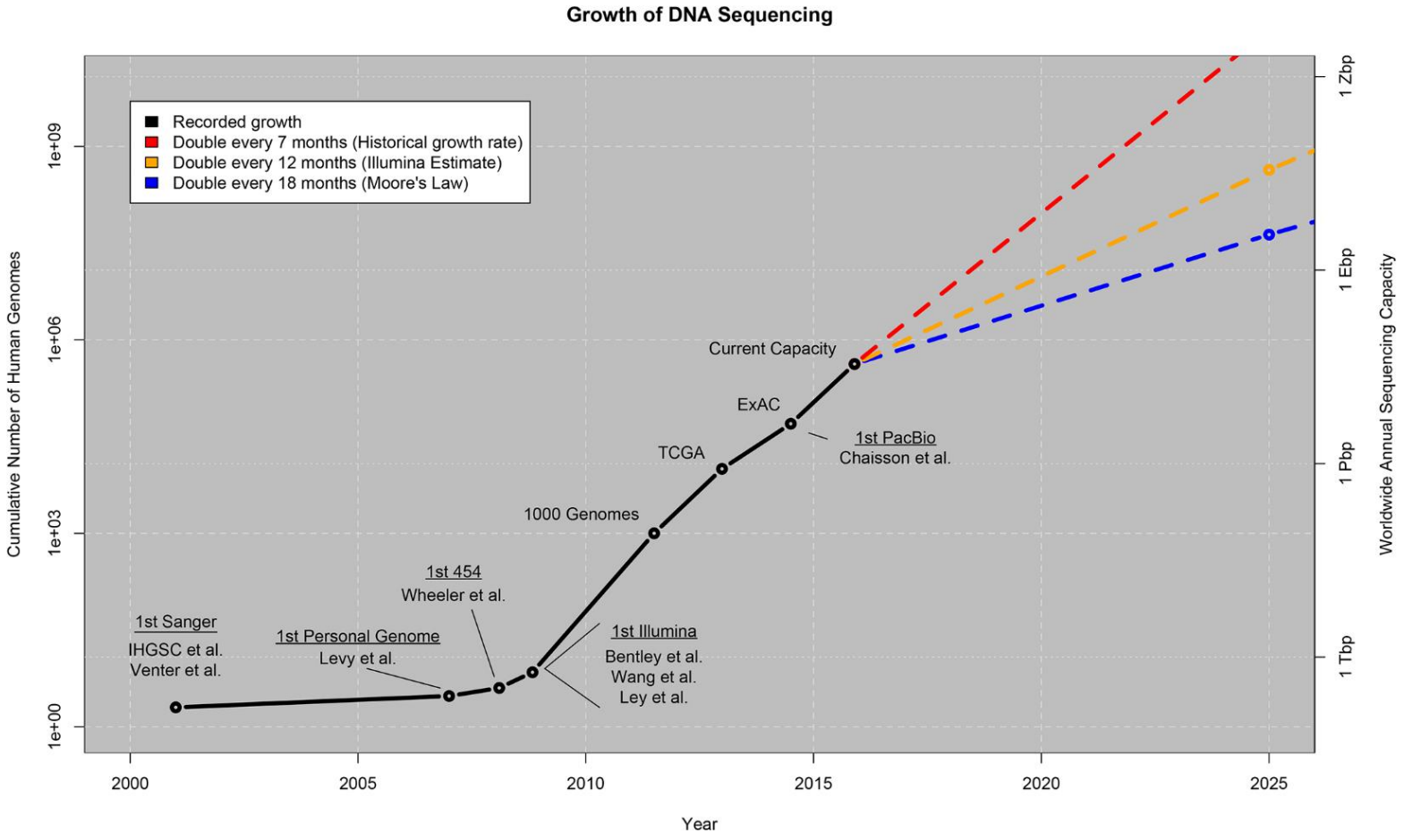
1:17 PM - 26 Oct 2016 from Manhattan, NY

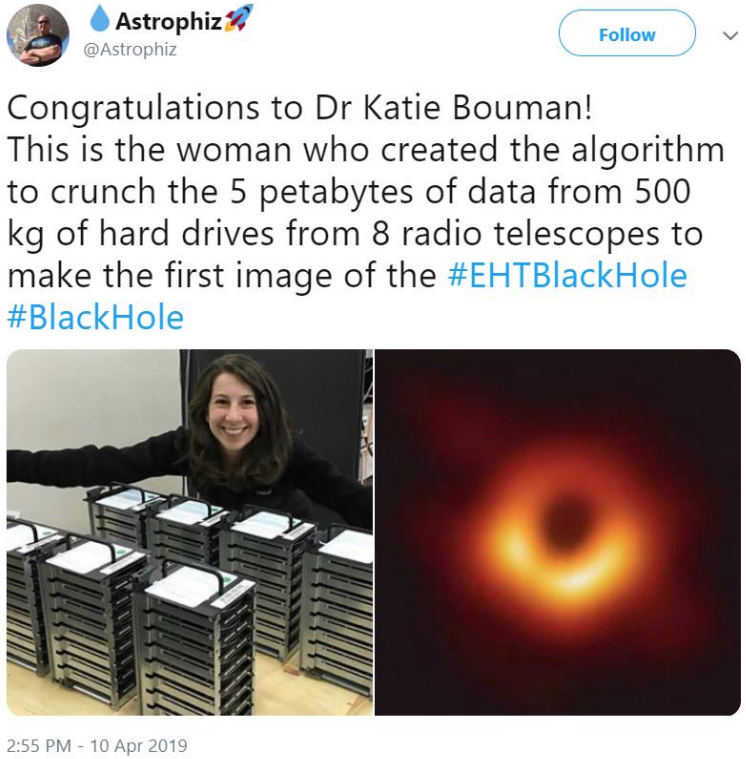
Who benefits most from reproducibility?



Where to begin…

The challenge





Stevens et. al., 2015, Plos Biology

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

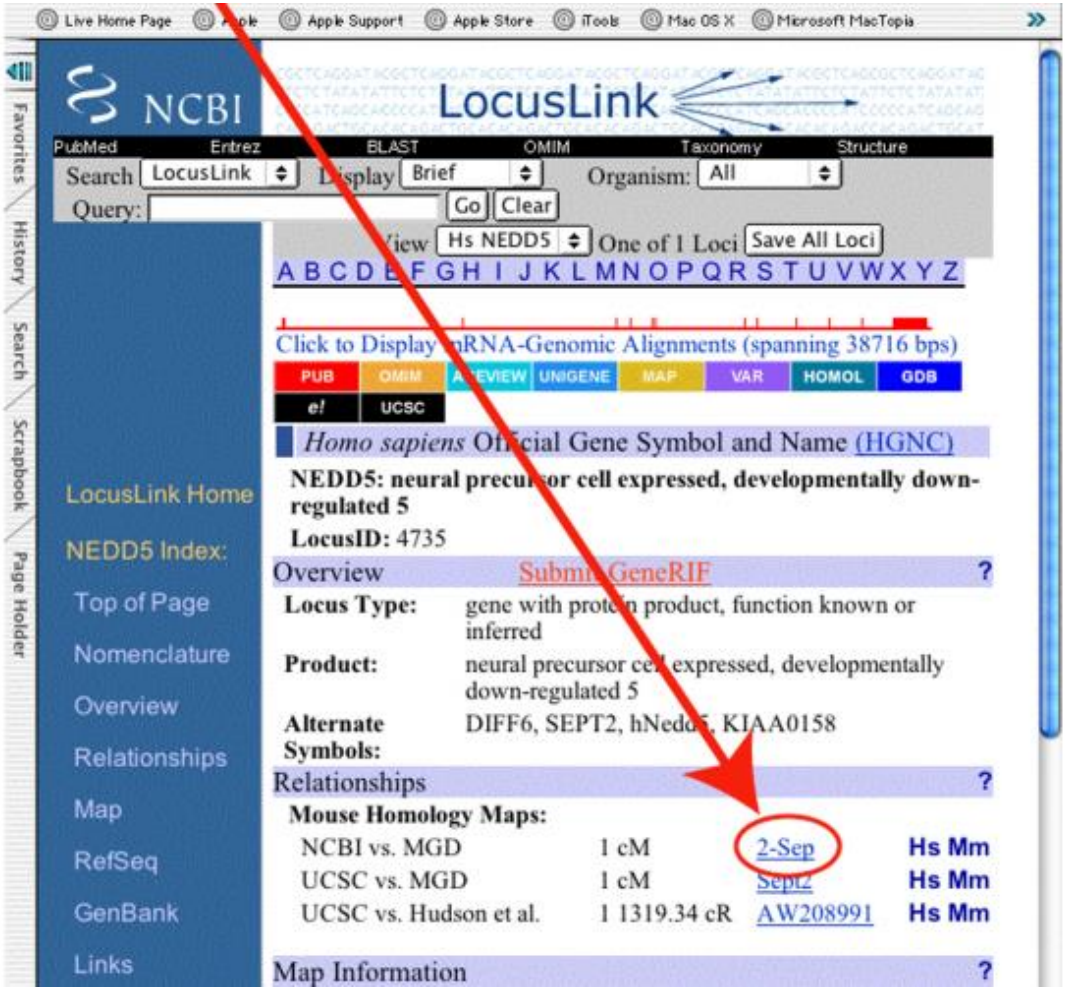
Mark Ziemann[1], Yotam Eren[1,2] and Assam El-Osta[1,3*]

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-



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	my	filter	1	1.545	2.45	cos	2019/04/15	Brendan Palmer		

data | dictionary | values

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									
4	3	0	ptp									
5	4	0	ptp									
6	1	0	my									
7	2	0	my									
8	3	0	my									
9	4	0	my									
10	1	0	ca									
11	2	0	ca									
12	3	0	ca									
13	4	0	ca									
14	5	1	ptp									
15	6	1	ptp									
16	7	1	ptp									
17	8	1	ptp									
18	9	1	my									
19	10	1	my									
20	11	1	my									
21	12	1	my									
22	13	1	ca									
23	14	1	ca									
24	15	1	ca									
25	16	1	ca									
26	17	2	ptp									
27	18	2	ptp									
28	19	2	ptp									
29	20	2	ptp									
30	21	2	my									

data | dictionary

	A	B	C	D	E	F
1	field_name	data_type	data_format	example	standard_units	description
2	id	numeric	integer	23	NA	Unique identifier applied to each observation
3	week_no	numeric	integer	1	NA	Week number, 1 = 7 days exposure, 2 = 14 days exposure
4	filter_name	character	NA	my	NA	3 filter types; 'ptp' = polytunnel plastic blocks all UV light
5	treatment	character	NA	filter	NA	Presence or absence of a filter at the time of sampling
6	replicate_no	numeric	integer	1	NA	The number of replicates in each treatement
7	flavonoids	numeric	double	0.3421	parts per million (ppm)	Leaf disc taken from the tip of the most mature leaf at th
8	biomass	numeric	double		gram (g)	Above ground biomass on the day of harvest
9	variety	character	NA	cos	NA	3 commerical varieties of red lettuce used; 'cos' = Cos Di
10	date	date	YYYY/MM/DD	2019/06/28	ISO 8601	Experiment date
11	investigator	character	Firstname Lastname	Aoife Coffey	NA	Primary researcher who performed the experiment
12						
13						
14						
15						
16						
17						
18						
19						
20						
21						
22						
23						
24						
25						
26						
27						
28						
29						
30						

data | dictionary | values

Less stress, more success

	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							
4	3	0	ptp							
5	4	0	ptp							
6	1	0	my							
7	2	0	my							
8	3	0	my							
9	4	0	my							
10	1	0	ca							
11	2	0	ca							
12	3	0	ca							
13	4	0	ca							
14	5	1	ptp							
15	6	1	ptp							
16	7	1	ptp							
17	8	1	ptp							
18	9	1	my							
19	10	1	my							
20	11	1	my							
21	12	1	my							
22	13	1	ca							
23	14	1	ca							
24	15	1	ca							
25	16	1	ca							
26	17	2	ptp							
27	18	2	ptp							
28	19	2	ptp							
29	20	2	ptp							
30	21	2	my							

data | dictionary

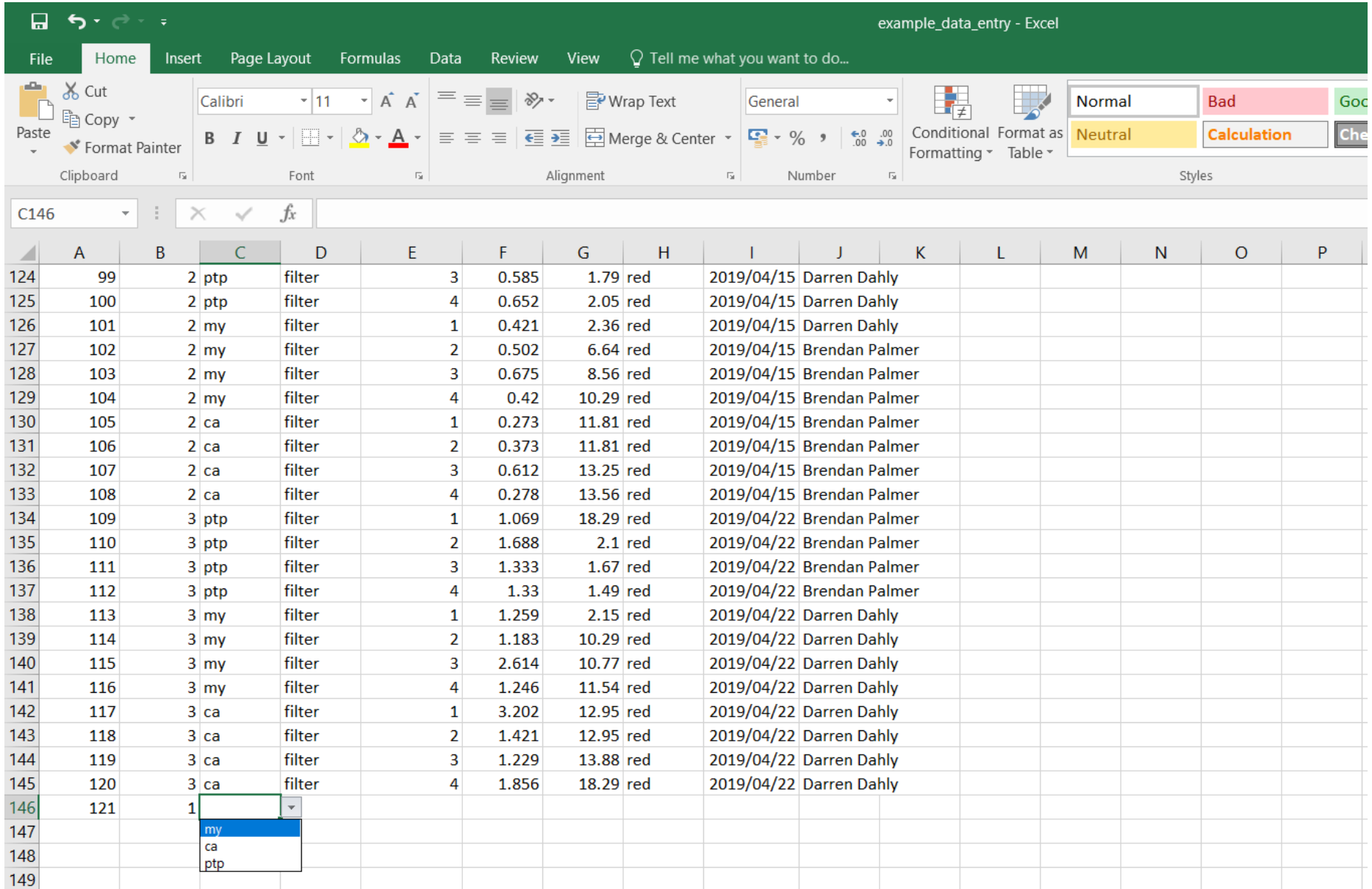
	field_name	data_type	data_format	example	standard_units	description
2	id	numeric	integer	23	NA	Unique identifier applied to each observation
3	week_no	numeric	integer			
4	filter_name	character	NA	n		
5	treatment	character	NA	fi		
6	replicate_no	numeric	integer			
7	flavonoids	numeric	double			
8	biomass	numeric	double			
9	variety	character	NA	c		
10	date	date	YYYY/MM/DD			
11	investigator	character	Firstname Lastname	A		

data | dictionary | values

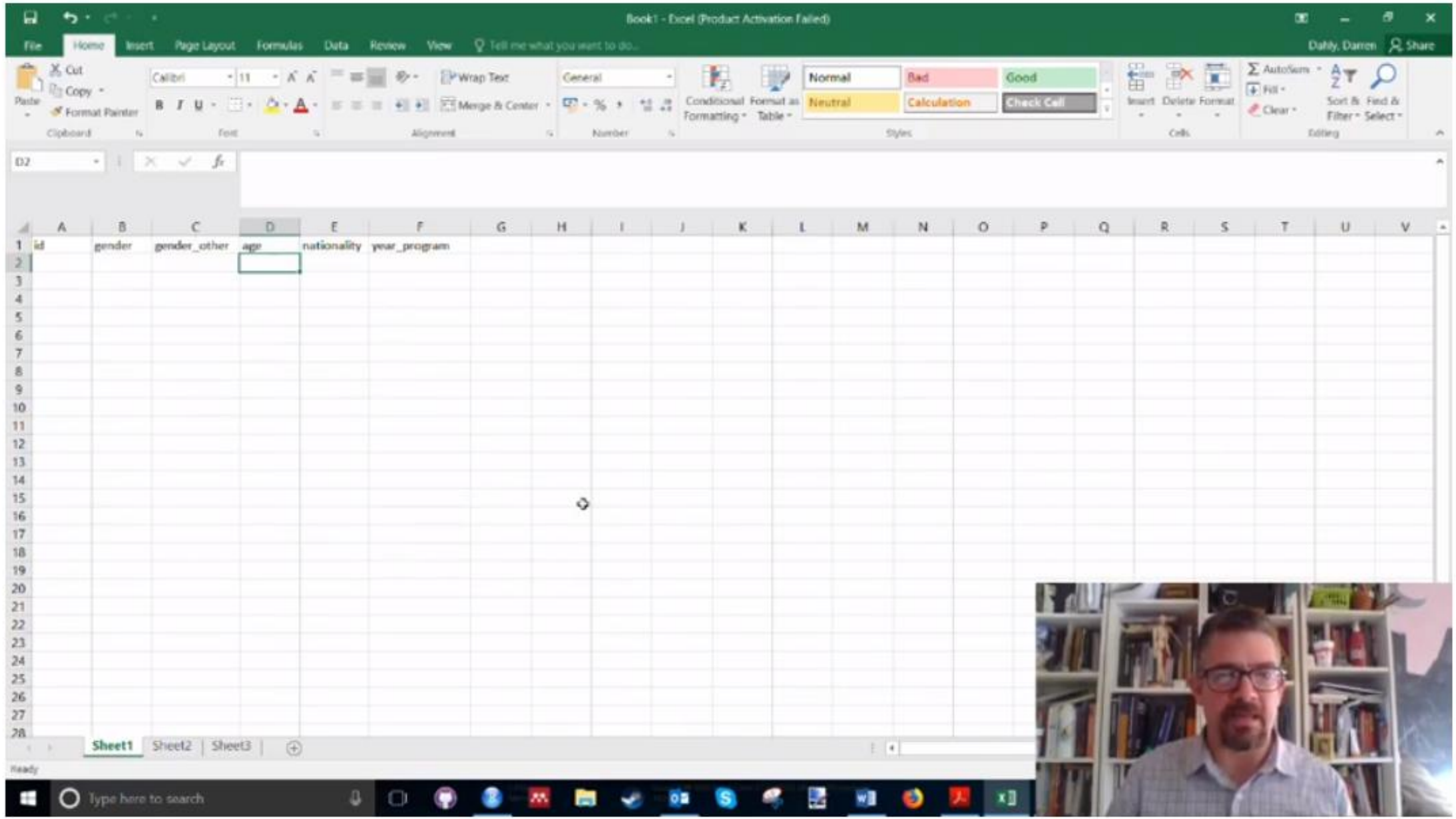
	id	week_no	filter_name	treatment	replicate_no	flavonoids	biomass	variety	date	investigator
2		0	my	filter	1			cos		Brendan Palmer
3		1	ca	no_filter	2			sky		Darren Dahly
4		2	ptp		3			red		
5		3			4					

data | dictionary | values

Less stress, more success



	A	B	C	D	E	F	G	H	I	J
124	99	2	ptp	filter	3	0.585	1.79	red	2019/04/15	Darren Dahly
125	100	2	ptp	filter	4	0.652	2.05	red	2019/04/15	Darren Dahly
126	101	2	my	filter	1	0.421	2.36	red	2019/04/15	Darren Dahly
127	102	2	my	filter	2	0.502	6.64	red	2019/04/15	Brendan Palmer
128	103	2	my	filter	3	0.675	8.56	red	2019/04/15	Brendan Palmer
129	104	2	my	filter	4	0.42	10.29	red	2019/04/15	Brendan Palmer
130	105	2	ca	filter	1	0.273	11.81	red	2019/04/15	Brendan Palmer
131	106	2	ca	filter	2	0.373	11.81	red	2019/04/15	Brendan Palmer
132	107	2	ca	filter	3	0.612	13.25	red	2019/04/15	Brendan Palmer
133	108	2	ca	filter	4	0.278	13.56	red	2019/04/15	Brendan Palmer
134	109	3	ptp	filter	1	1.069	18.29	red	2019/04/22	Brendan Palmer
135	110	3	ptp	filter	2	1.688	2.1	red	2019/04/22	Brendan Palmer
136	111	3	ptp	filter	3	1.333	1.67	red	2019/04/22	Brendan Palmer
137	112	3	ptp	filter	4	1.33	1.49	red	2019/04/22	Brendan Palmer
138	113	3	my	filter	1	1.259	2.15	red	2019/04/22	Darren Dahly
139	114	3	my	filter	2	1.183	10.29	red	2019/04/22	Darren Dahly
140	115	3	my	filter	3	2.614	10.77	red	2019/04/22	Darren Dahly
141	116	3	my	filter	4	1.246	11.54	red	2019/04/22	Darren Dahly
142	117	3	ca	filter	1	3.202	12.95	red	2019/04/22	Darren Dahly
143	118	3	ca	filter	2	1.421	12.95	red	2019/04/22	Darren Dahly
144	119	3	ca	filter	3	1.229	13.88	red	2019/04/22	Darren Dahly
145	120	3	ca	filter	4	1.856	18.29	red	2019/04/22	Darren Dahly
146	121	1								
147										
148										
149										



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

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

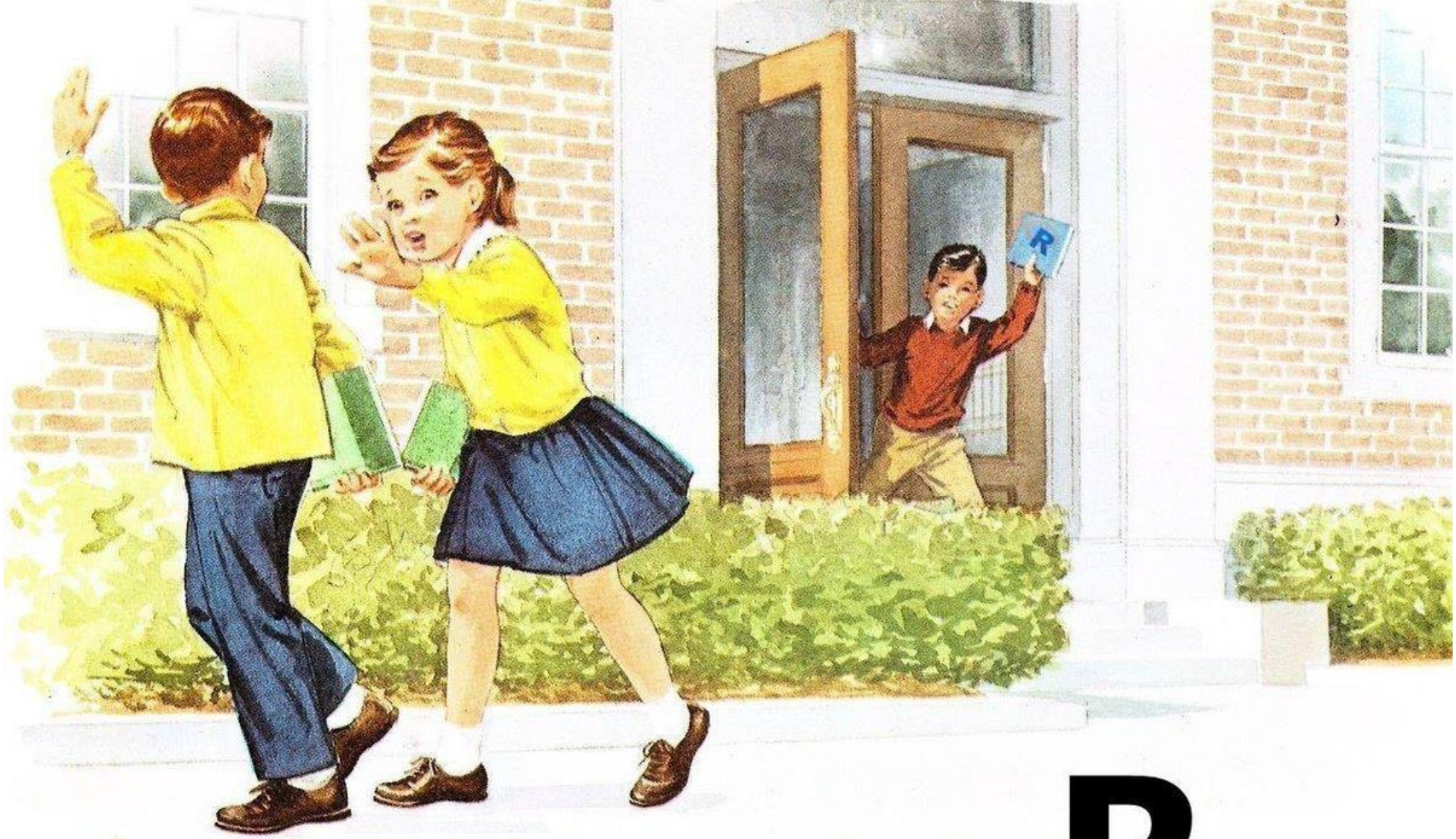
A: Define a project structure

B: Set a naming convention

C: Use scripted workflows

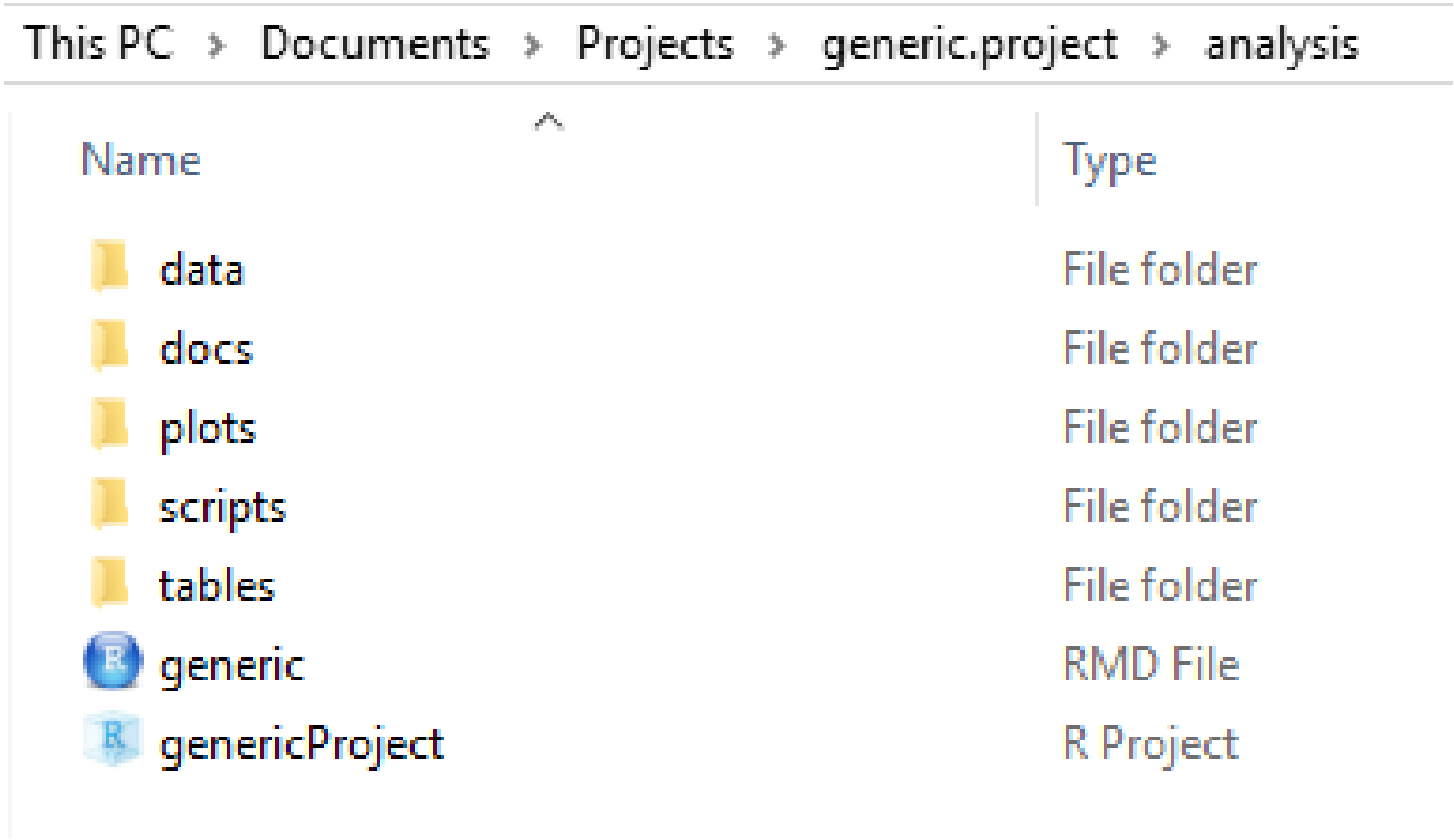
D: Digital reports

Reproducible research



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

A: Define a generic project structure



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:

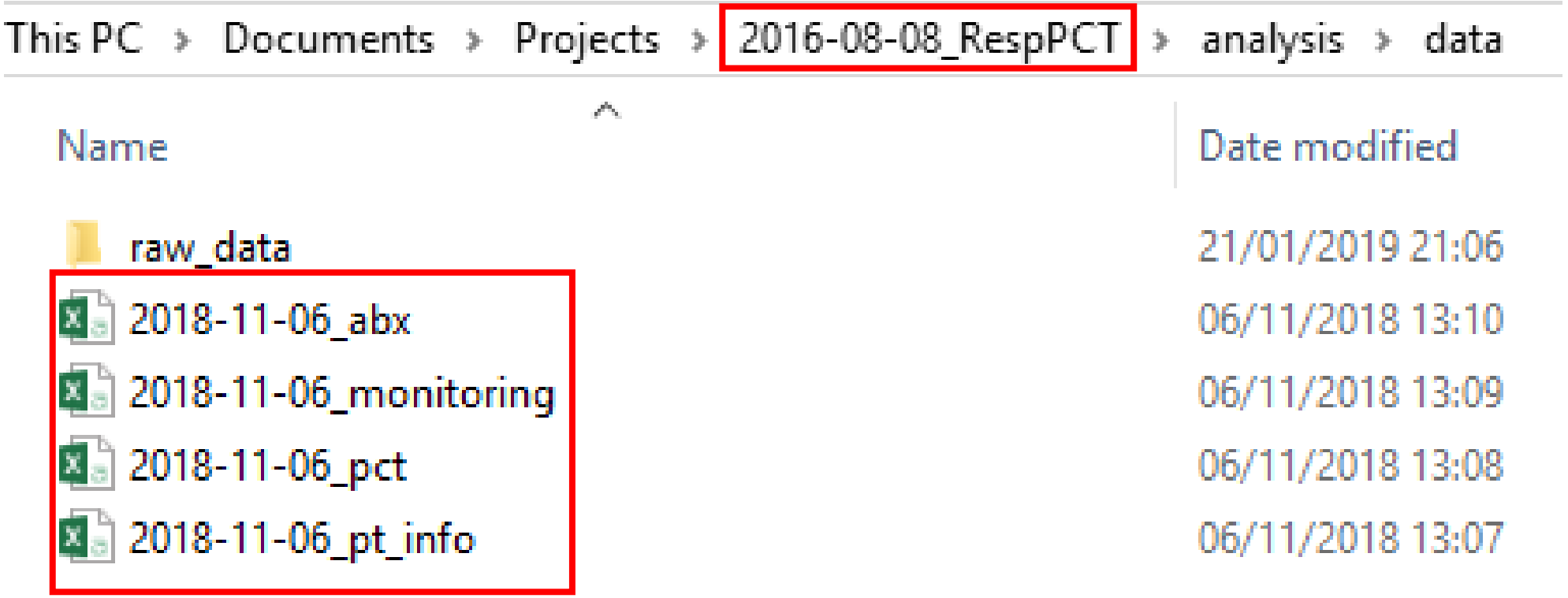
Separate with underscores ("_")

- Avoid punctuation
- Remove case-sensitivity

```
01_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
```

Image from Data Carpentry webpage

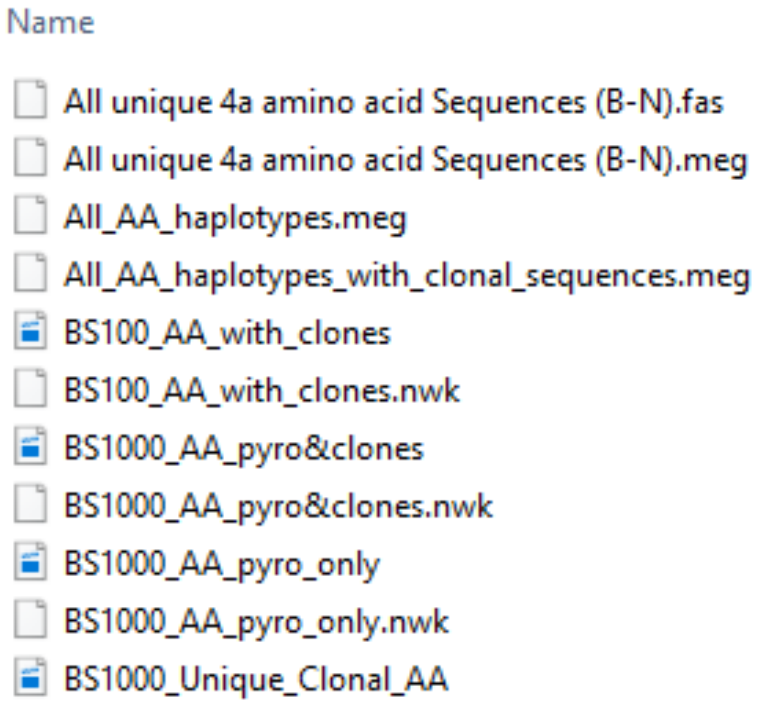
B: Give your files and folders informative names



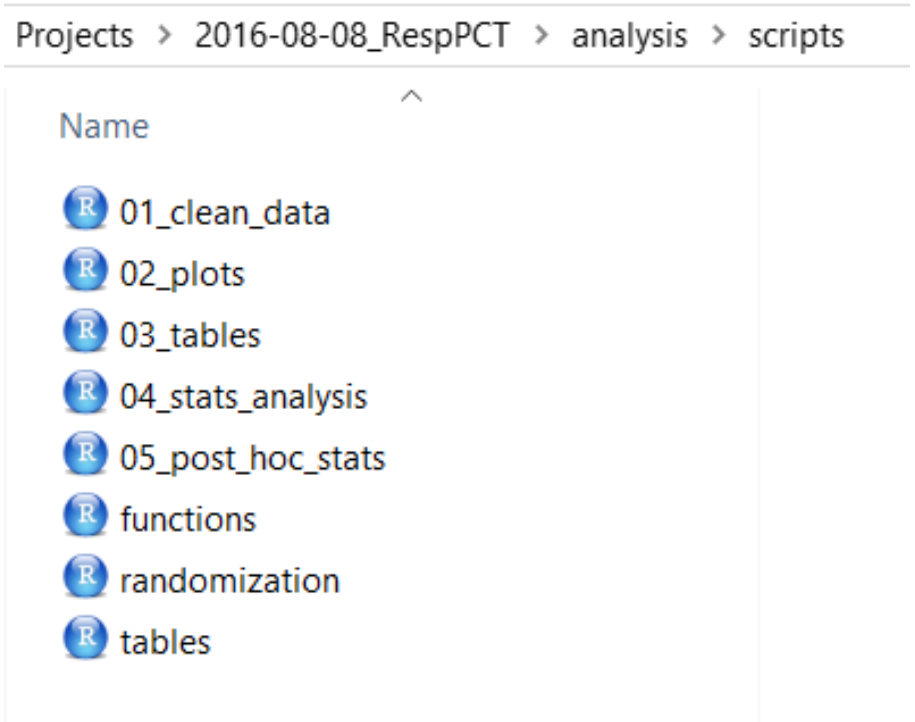
Everything in its right place

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

NO



Yes



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

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

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

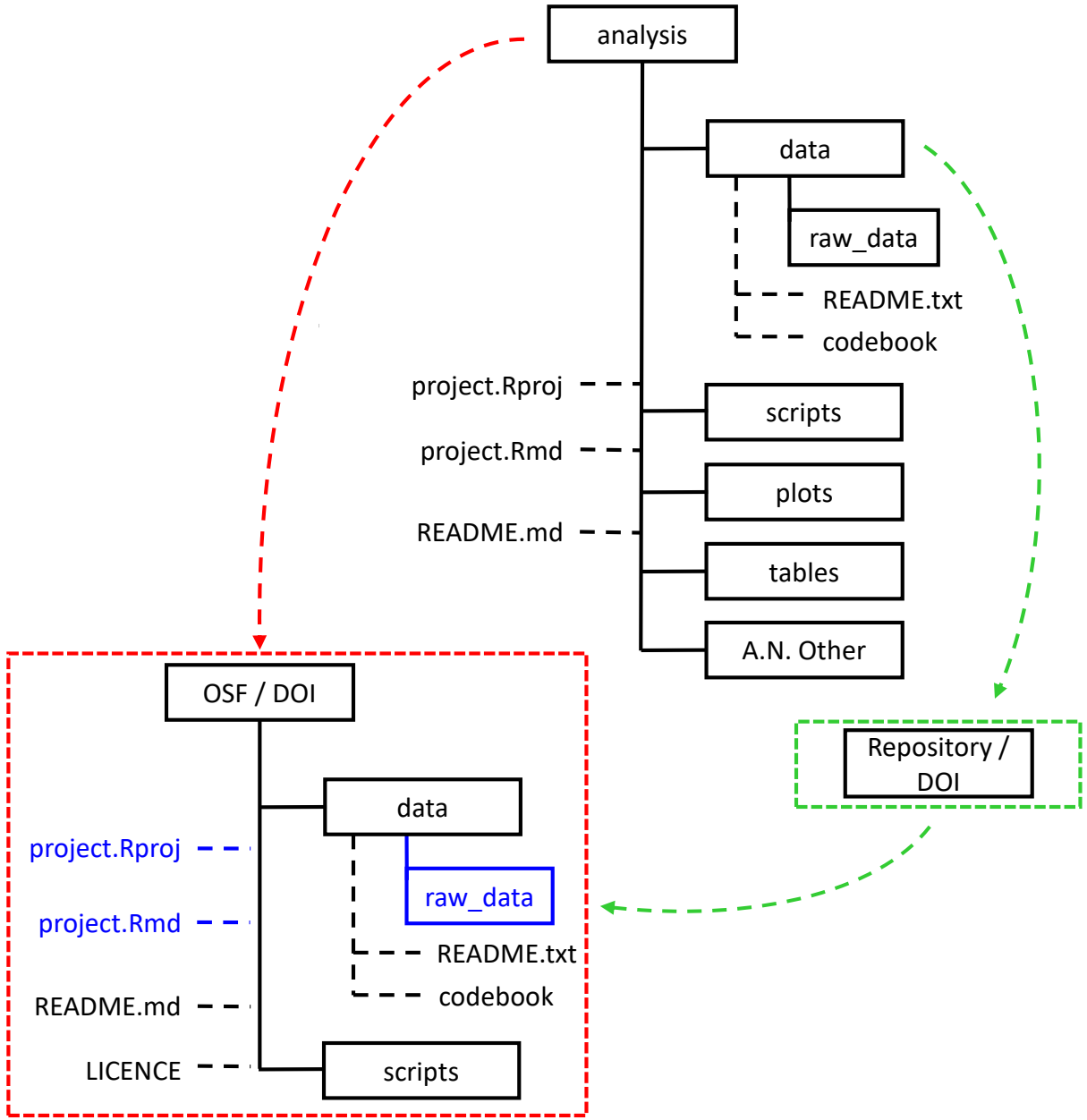
````
We have data about `r nrow(diamonds)`
diamonds. Only
`r nrow(diamonds) - nrow(smaller)` are
larger than
2.5 carats. The distribution of the
remainder is shown below:
````

Chunks of code

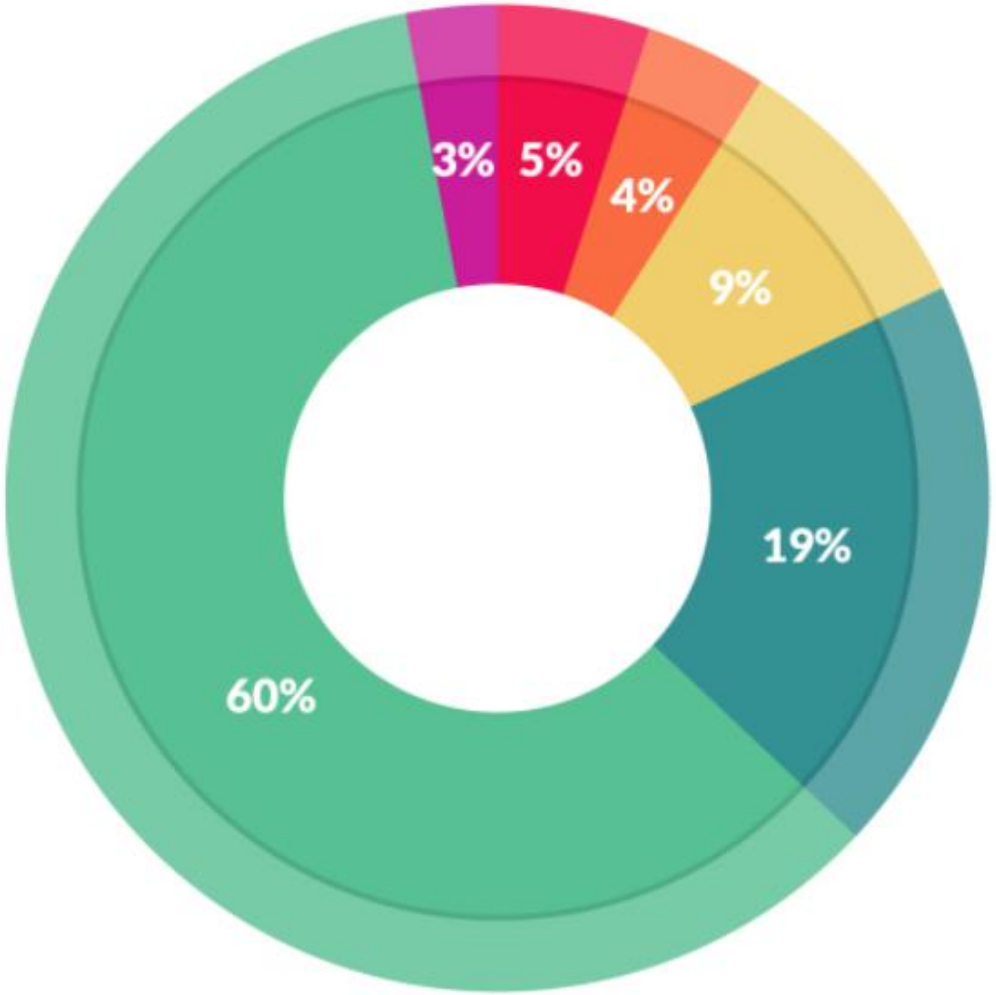
````
```{r, echo = FALSE}
smaller %>%
ggplot(aes(carat)) +
geom_freqpoly(binwidth = 0.01)
```
````

Install the Chrome plugin PubPeer

What does this allow us to do?



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%*

CrowdFlower (2016) Data Science Report

The best time to plant a tree was 20 years ago

The second best time is now