Digital badge

- Reproducible research

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

CRF-C
HRB Clinical Research Facility Cork

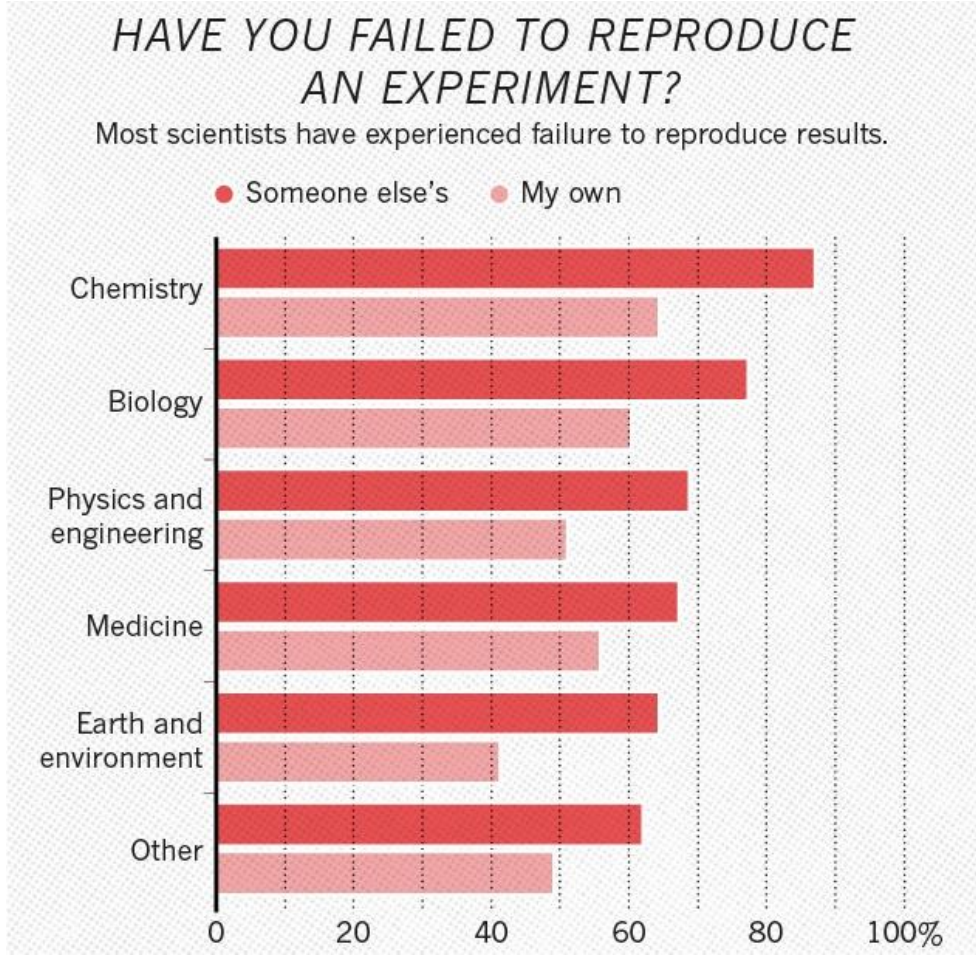
University College Cork
2019
Responsible Conduct of Research
Coláiste na hOllscoile Corcaigh

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

1,500 scientists lift the lid on reproducibility

Survey sheds light on the 'crisis' rocking research.

Monya Baker



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

Reproducible or replicable

		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

Reproducible or replicable

		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

Reproducible or replicable

		Data	
		Same	Different
Analysis	Same	Replicable	Reproducible
	Different	Robust	Generalisable

Some cautionary tales

The Atlantic Popular Latest Sections

SCIENCE

A Waste of 1,000 Research Papers

Decades of early research on the genetics of depression were built on nonexistent foundations. How did that happen?

ED YONG MAY 17, 2019

Am J Psychiatry. 2019 May 1;176(5):376-387. doi: 10.1176/appi.ajp.2018.18070881. Epub 2019 Mar 8.

No Support for Historical Candidate Gene or Candidate Gene-by-Interaction Hypotheses for Major Depression Across Multiple Large Samples.

Border R[1], Johnson EC[1], Evans LM[1], Smolen A[1], Berley N[1], Sullivan PF[1], Keller MC[1].

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.

Who benefits most from reproducibility?

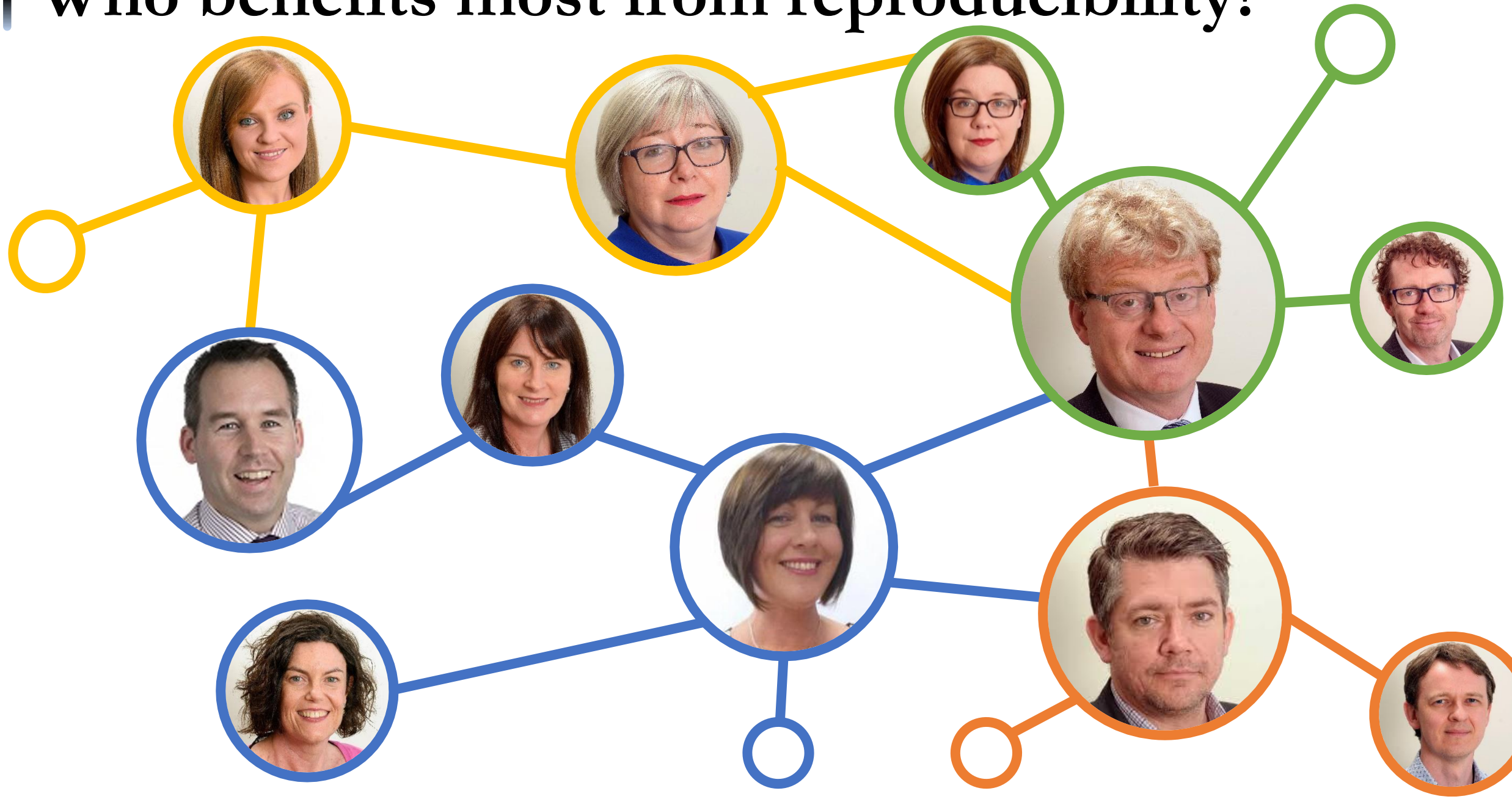
Casey Greene
@GreeneScientist

Follow

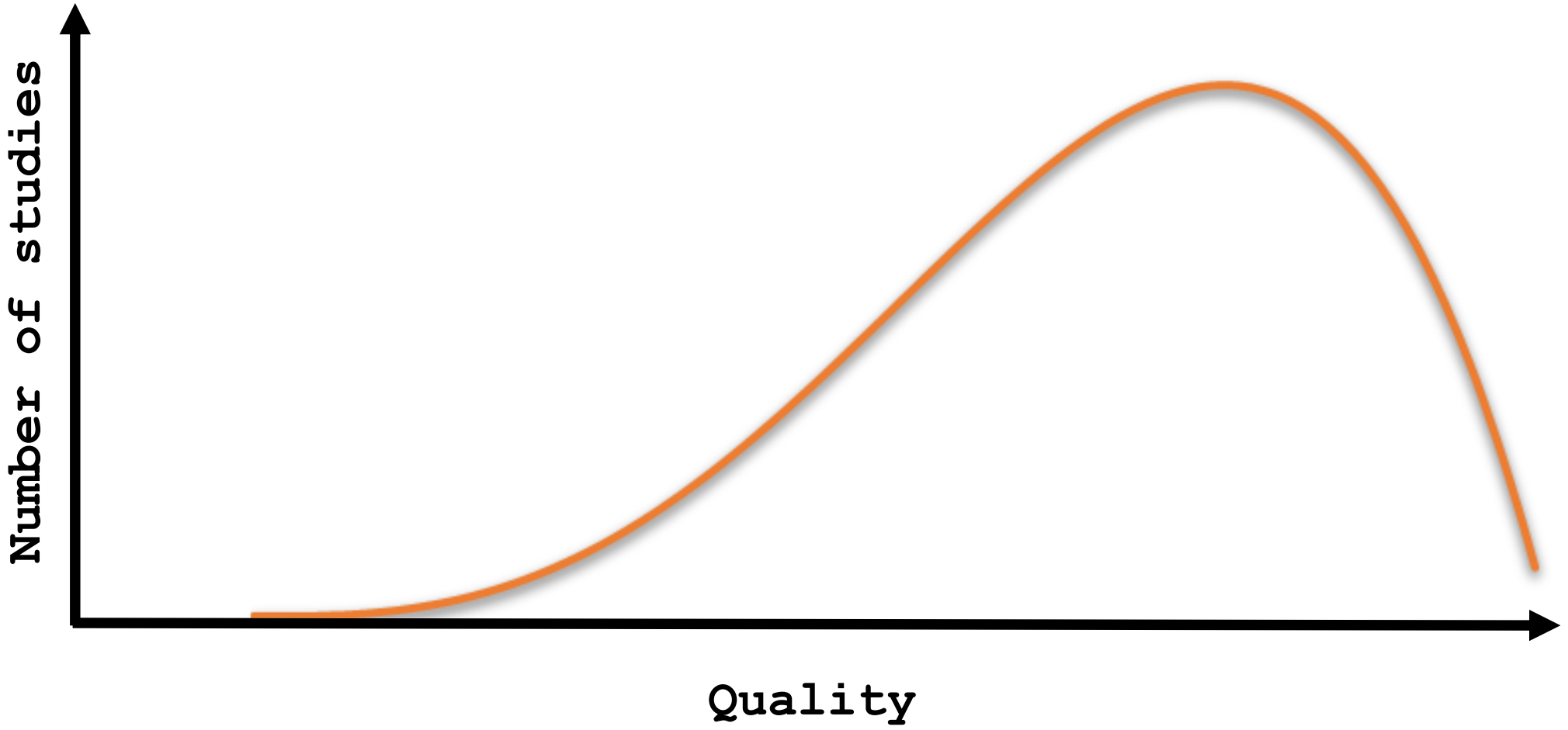
Reproducibility is important because the you of 3 months ago is terrible at answering email! - @tracykteal at #2016dssummit

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

Who benefits most from reproducibility?



Today

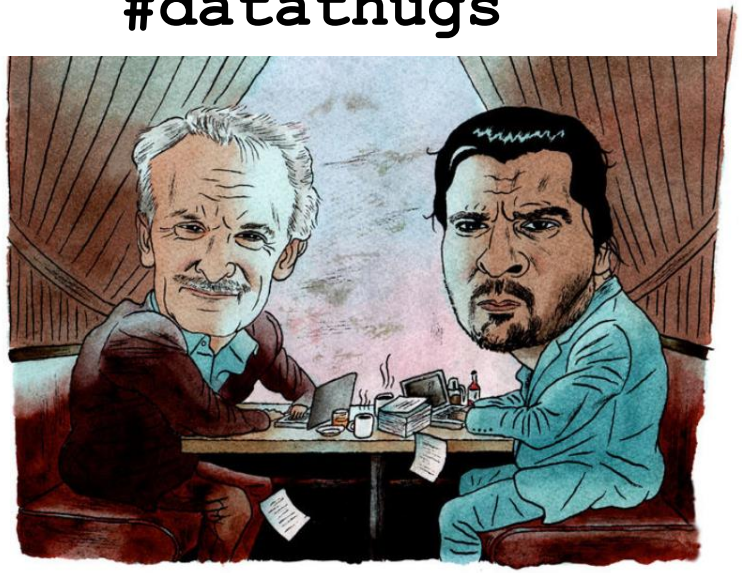


Past failings

"In short, peer review misses all the hard stuff, and a worrying amount of the easy stuff"

James Heathers,
Northwestern University

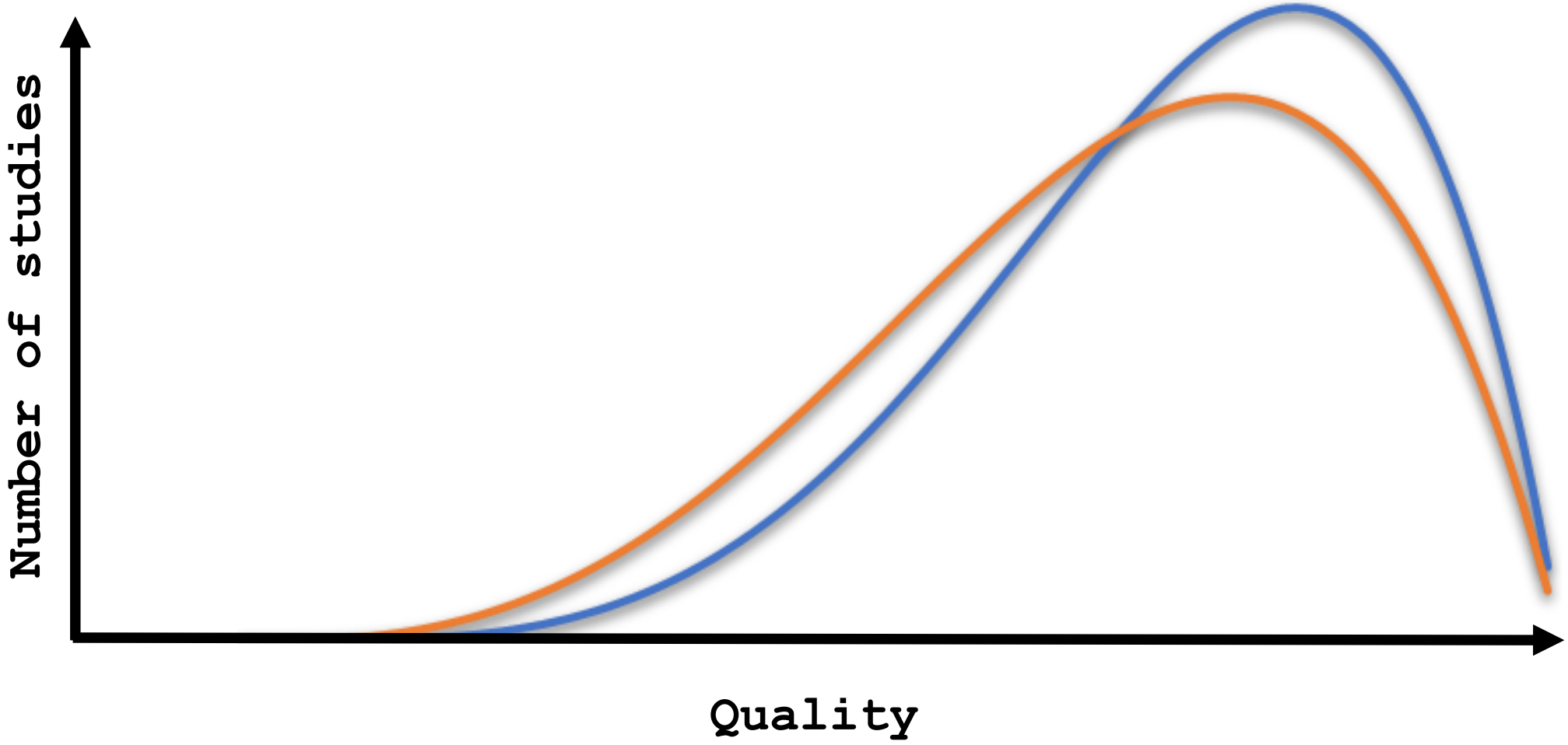
#datathugs



Brian Wansink: The grad student who never said no

"Every day we would scratch our heads, ask "Why," and come up with another way to reanalyze the data with yet another set of plausible hypotheses. Eventually we started discovering solutions"

Tomorrow



Where to begin…

Fundamental problem

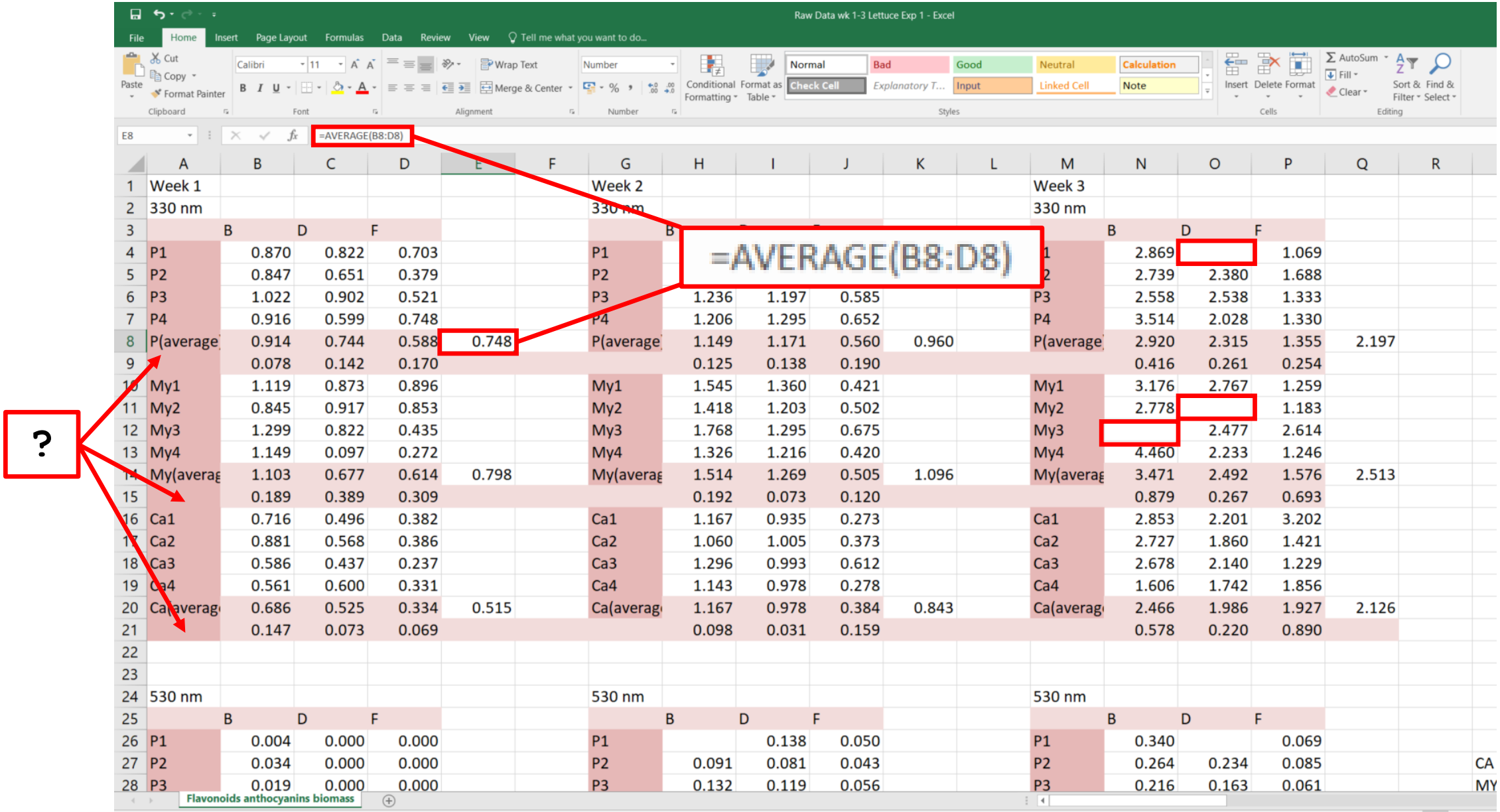
I'm not in the office at the moment. Send any work to be translated

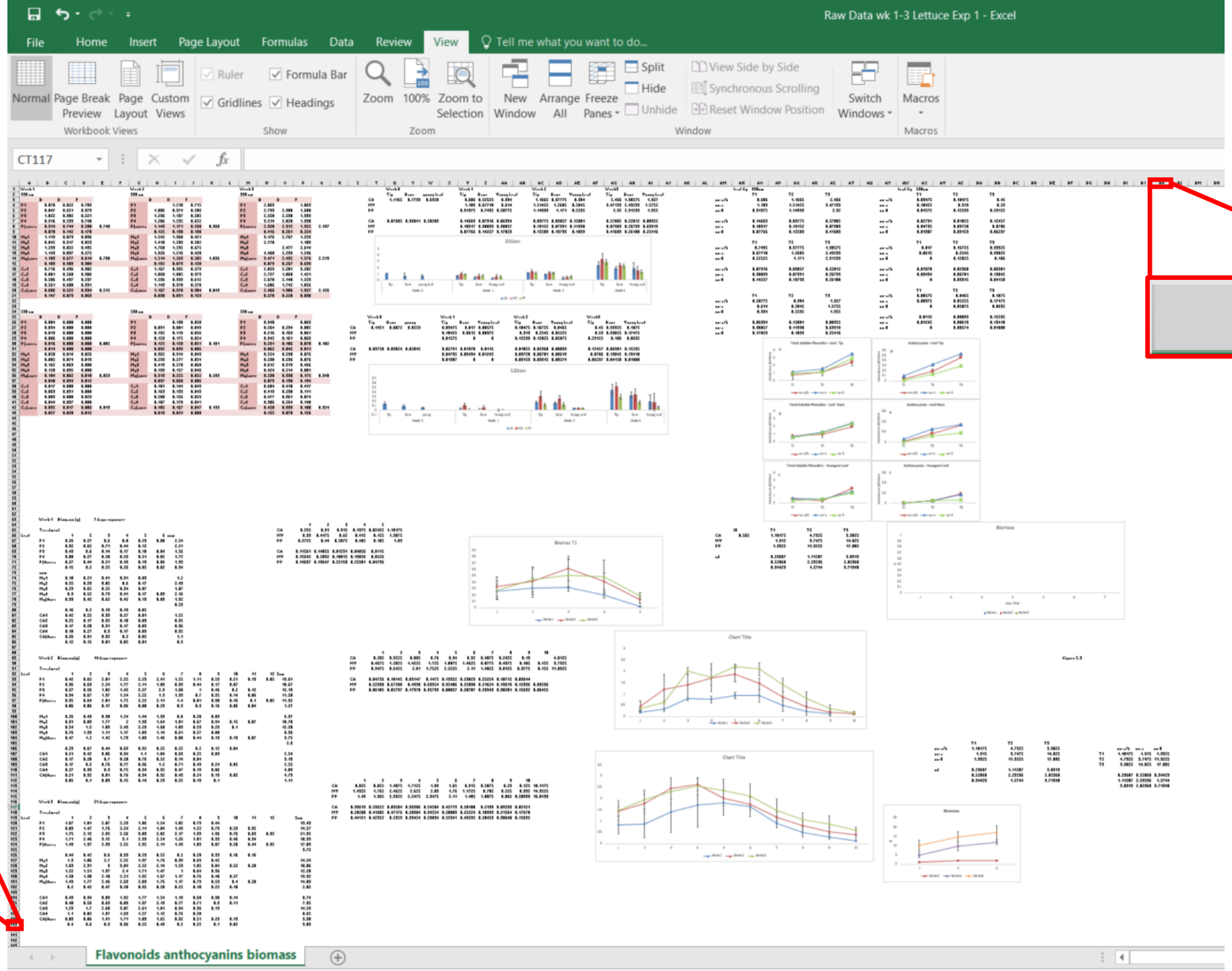
Our real life experiment

- UV light has potential to change the secondary metabolite composition (colour) of bronze/red lettuce

- Experimental setup:
  - 3 lettuce varieties
  - 3 UV filter conditions
  - 3 week duration

Real data comes with real problems




Raw Data wk 1-3 Lettuce Exp 1 - Excel
File
Home
Insert
Page Layout
Formulas
Data
Review
View
Tell me what you want to do...
Normal
Page Break Preview
Page Layout
Custom Views
Workbook Views
Ruler
Gridlines
Formula Bar
Headings
Show
Zoom
100%
Zoom to Selection
Zoom
New Window
Arrange All
Freeze Panes
Split
Hide
Unhide
View Side by Side
Synchronous Scrolling
Reset Window Position
Window
Switch Windows
Macros
CT117
Flavonoids anthocyanins biomass

BK

139

Raw Data wk 1-3 Lettuce Exp 1 - Excel

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

Normal Page Break Preview Page Layout Custom Views — Workbook Views

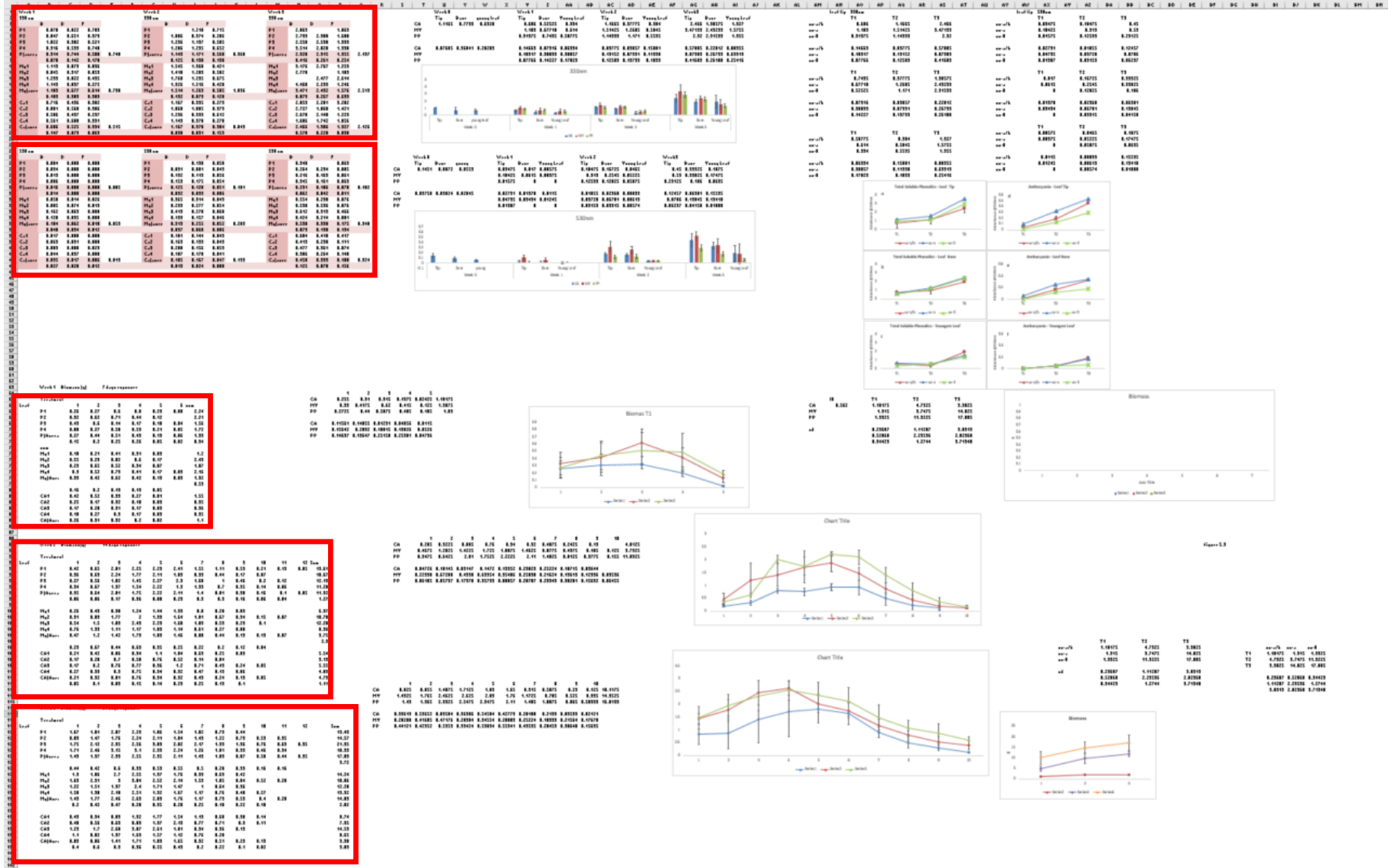
Ruler Formula Bar Gridlines Headings — Show

Zoom 100% Zoom to Selection — Zoom

New Window Arrange All Freeze Panes Split Hide Unhide View Side by Side Synchronous Scrolling Reset Window Position Switch Windows — Window

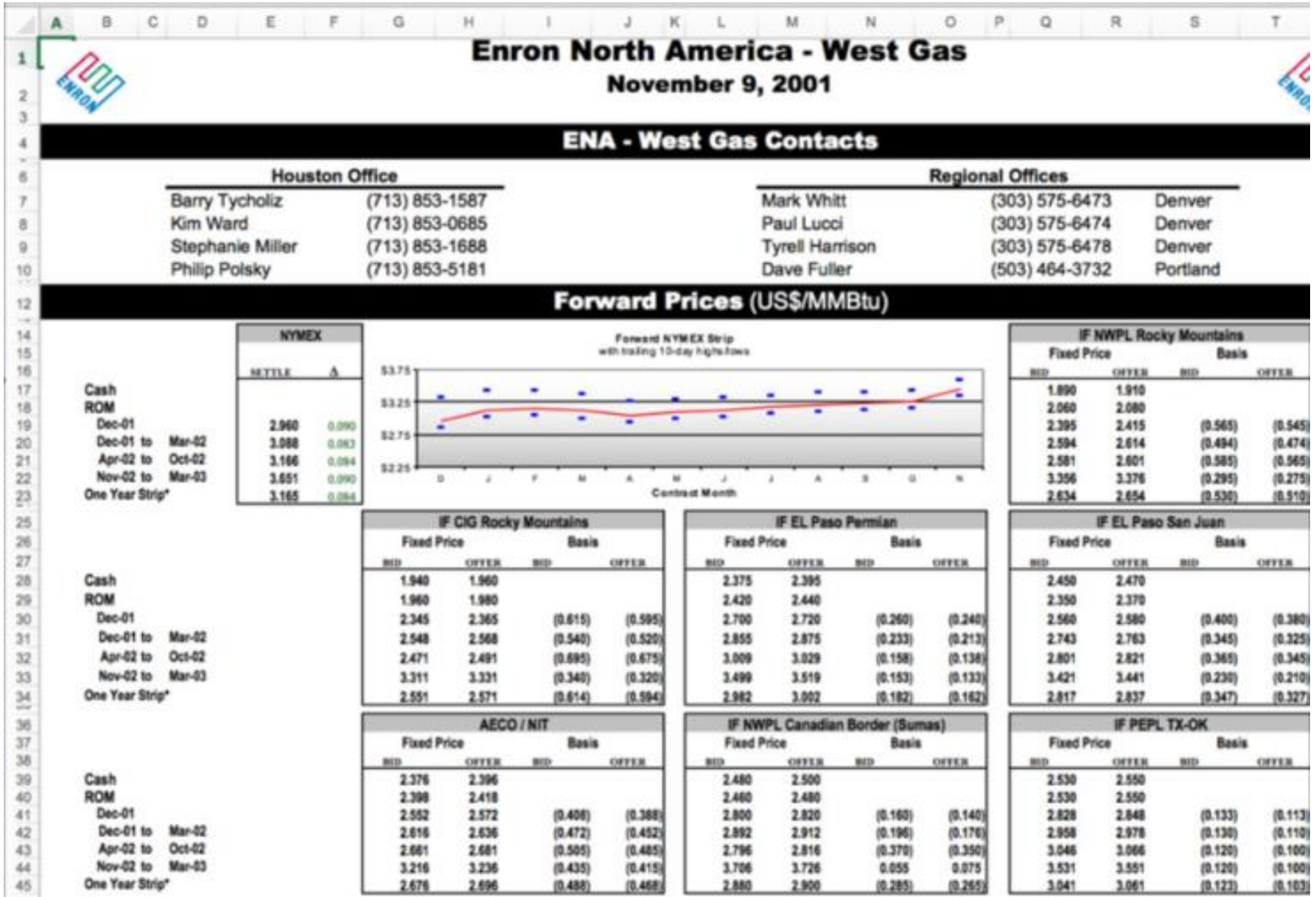
Macros — Macros

CT117



Flavonoids anthocyanins biomass

This is a big problem



Enron North America - West Gas

November 9, 2001

ENA - West Gas Contacts

Houston Office		Regional Offices		
Barry Tycholiz	(713) 853-1587	Mark Whitt	(303) 575-6473	Denver
Kim Ward	(713) 853-0685	Paul Lucci	(303) 575-6474	Denver
Stephanie Miller	(713) 853-1688	Tyrell Harrison	(303) 575-6478	Denver
Philip Polsky	(713) 853-5181	Dave Fuller	(503) 464-3732	Portland

Forward Prices (US$/MMBtu)

	NYMEX SETTLE	NYMEX Δ
Cash		
ROM		
Dec-01	2.960	0.090
Dec-01 to Mar-02	3.088	0.083
Apr-02 to Oct-02	3.166	0.084
Nov-02 to Mar-03	3.651	0.090
One Year Strip*	3.165	0.084

Forward NYMEX Strip
with trailing 10-day highs/lows

$3.75
$3.25
$2.75
$2.25

D J F M A M J J A S O N

Contract Month

	IF NWPL Rocky Mountains			
	Fixed Price		Basis	
	BID	OFFER	BID	OFFER
Cash	1.890	1.910		
ROM	2.060	2.080		
Dec-01	2.395	2.415	(0.565)	(0.545)
Dec-01 to Mar-02	2.594	2.614	(0.494)	(0.474)
Apr-02 to Oct-02	2.581	2.601	(0.585)	(0.565)
Nov-02 to Mar-03	3.356	3.376	(0.295)	(0.275)
One Year Strip*	2.634	2.654	(0.530)	(0.510)

	IF CIG Rocky Mountains			
	Fixed Price		Basis	
	BID	OFFER	BID	OFFER
Cash	1.940	1.960		
ROM	1.960	1.980		
Dec-01	2.345	2.365	(0.615)	(0.595)
Dec-01 to Mar-02	2.548	2.568	(0.540)	(0.520)
Apr-02 to Oct-02	2.471	2.491	(0.695)	(0.675)
Nov-02 to Mar-03	3.311	3.331	(0.340)	(0.320)
One Year Strip*	2.551	2.571	(0.614)	(0.594)

	IF EL Paso Permian			
	Fixed Price		Basis	
	BID	OFFER	BID	OFFER
Cash	2.375	2.395		
ROM	2.420	2.440		
Dec-01	2.700	2.720	(0.260)	(0.240)
Dec-01 to Mar-02	2.855	2.875	(0.233)	(0.213)
Apr-02 to Oct-02	3.009	3.029	(0.158)	(0.138)
Nov-02 to Mar-03	3.499	3.519	(0.153)	(0.133)
One Year Strip*	2.982	3.002	(0.182)	(0.162)

	IF EL Paso San Juan			
	Fixed Price		Basis	
	BID	OFFER	BID	OFFER
Cash	2.450	2.470		
ROM	2.350	2.370		
Dec-01	2.560	2.580	(0.400)	(0.380)
Dec-01 to Mar-02	2.743	2.763	(0.345)	(0.325)
Apr-02 to Oct-02	2.801	2.821	(0.365)	(0.345)
Nov-02 to Mar-03	3.421	3.441	(0.230)	(0.210)
One Year Strip*	2.817	2.837	(0.347)	(0.327)

	AECO / NIT			
	Fixed Price		Basis	
	BID	OFFER	BID	OFFER
Cash	2.376	2.396		
ROM	2.398	2.418		
Dec-01	2.552	2.572	(0.408)	(0.388)
Dec-01 to Mar-02	2.616	2.636	(0.472)	(0.452)
Apr-02 to Oct-02	2.661	2.681	(0.505)	(0.485)
Nov-02 to Mar-03	3.216	3.236	(0.435)	(0.415)
One Year Strip*	2.676	2.696	(0.488)	(0.468)

	IF NWPL Canadian Border (Sumas)			
	Fixed Price		Basis	
	BID	OFFER	BID	OFFER
Cash	2.480	2.500		
ROM	2.460	2.480		
Dec-01	2.800	2.820	(0.160)	(0.140)
Dec-01 to Mar-02	2.892	2.912	(0.196)	(0.176)
Apr-02 to Oct-02	2.796	2.816	(0.370)	(0.350)
Nov-02 to Mar-03	3.706	3.726	0.055	0.075
One Year Strip*	2.880	2.900	(0.285)	(0.265)

	IF PEPL TX-OK			
	Fixed Price		Basis	
	BID	OFFER	BID	OFFER
Cash	2.530	2.550		
ROM	2.530	2.550		
Dec-01	2.828	2.848	(0.133)	(0.113)
Dec-01 to Mar-02	2.958	2.978	(0.130)	(0.110)
Apr-02 to Oct-02	3.046	3.066	(0.120)	(0.100)
Nov-02 to Mar-03	3.531	3.551	(0.120)	(0.100)
One Year Strip*	3.041	3.061	(0.123)	(0.103)

Take small steps to big changes

THE AMERICAN STATISTICIAN
2018, VOL. 72, NO. 1, 2–10
https://doi.org/10.1080/00031305.2017.1375989

Data Organization in Spreadsheets

Karl W. Broman[a] and Kara H. Woo[b]

[a]Department of Biostatistics & Medical Informatics, University of Wisconsin-Madison, Madison, WI; [b]Information School, University of Washington, Seattle, WA

ABSTRACT

Spreadsheets are widely used software tools for data entry, storage, analysis, and visualization. Focusing on the data entry and storage aspects, this article offers practical recommendations for organizing spreadsheet data to reduce errors and ease later analyses. The basic principles are: be consistent, write dates like YYYY-MM-DD, do not leave any cells empty, put just one thing in a cell, organize the data as a single rectangle (with subjects as rows and variables as columns, and with a single header row), create a data dictionary, do not include calculations in the raw data files, do not use font color or highlighting as data, choose good names for things, make backups, use data validation to avoid data entry errors, and save the data in plain text files.

ARTICLE HISTORY
Received June 2017
Revised August 2017

KEYWORDS
Data management; Data organization; Microsoft Excel; Spreadsheets

Less stress, more success

	A	B	C	D	E	F	G	H	I	J	K
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	5	0	my	nofilter	1	1.061	0.39	cos	2019/04/01	Brendan Palmer	
7	6	0	my	nofilter	2	1.1805	0.42	cos	2019/04/01	Brendan Palmer	
8	7	0	my	nofilter	3	1.0345	0.62	cos	2019/04/01	Brendan Palmer	
9	8	0	my	nofilter	4	1.094	0.63	cos	2019/04/01	Brendan Palmer	
10	9	0	ca	nofilter	1	1.061	0.39	cos	2019/04/01	Brendan Palmer	
11	10	0	ca	nofilter	2	1.1805	0.42	cos	2019/04/01	Brendan Palmer	
12	11	0	ca	nofilter	3	1.0345	0.62	cos	2019/04/01	Brendan Palmer	
13	12	0	ca	nofilter	4	1.094	0.63	cos	2019/04/01	Darren Dahly	
14	13	1	ptp	filter	1	0.87	0.76	cos	2019/04/08	Darren Dahly	
15	14	1	ptp	filter	2	0.847	0.95	cos	2019/04/08	Darren Dahly	
16	15	1	ptp	filter	3	1.022	0.95	cos	2019/04/08	Darren Dahly	
17	16	1	ptp	filter	4	0.916	0.95	cos	2019/04/08	Darren Dahly	
18	17	1	my	filter	1	1.119	1.55	cos	2019/04/08	Darren Dahly	
19	18	1	my	filter	2	0.845	3.16	cos	2019/04/08	Darren Dahly	
20	19	1	my	filter	3	1.299	4.9	cos	2019/04/08	Brendan Palmer	
21	20	1	my	filter	4	1.149	5.5	cos	2019/04/08	Brendan Palmer	
22	21	1	ca	filter	1	0.716	5.5	cos	2019/04/08	Brendan Palmer	
23	22	1	ca	filter	2	0.881	7.94	cos	2019/04/08	Brendan Palmer	
24	23	1	ca	filter	3	0.586	8.71	cos	2019/04/08	Brendan Palmer	
25	24	1	ca	filter	4	0.561	8.71	cos	2019/04/08	Brendan Palmer	
26	25	2	ptp	filter	1	0	14.45	cos	2019/04/15	Brendan Palmer	
27	26	2	ptp	filter	2	1.006	2.14	cos	2019/04/15	Brendan Palmer	
28	27	2	ptp	filter	3	1.236	1.86	cos	2019/04/15	Brendan Palmer	
29	28	2	ptp	filter	4	1.206	1.2	cos	2019/04/15	Brendan Palmer	
30	29	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
1	id	week_no	filter_nam	treatment	replicate_no	flavonoids	biomass	variety	date	investigator	
2	1	0	ptp	no							
3	2	0	ptp	no							
4	3	0	ptp	no							
5	4	0	ptp	no							
6	5	0	my	no							
7	6	0	my	no							
8	7	0	my	no							
9	8	0	my	no							
10	9	0	ca	no							
11	10	0	ca	no							
12	11	0	ca	no							
13	12	0	ca	no							
14	13	1	ptp	fil							
15	14	1	ptp	fil							
16	15	1	ptp	fil							
17	16	1	ptp	fil							
18	17	1	my	fil							
19	18	1	my	fil							
20	19	1	my	fil							
21	20	1	my	fil							
22	21	1	ca	fil							
23	22	1	ca	fil							
24	23	1	ca	fil							
25	24	1	ca	fil							
26	25	2	ptp	fil							
27	26	2	ptp	fil							
28	27	2	ptp	fil							
29	28	2	ptp	fil							
30	29	2	my	fil							

data | dictionary | v

	A	B	C	D	E	
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
1	0	ptp	no filter	1	1.061	0.29	cos	2019/04/01	Darren Dahly
2	0	ptp	no						
3	0	ptp	no						
4	0	ptp	no						
5	0	my	no						
6	0	my	no						
7	0	my	no						
8	0	my	no						
9	0	ca	no						
10	0	ca	no						
11	0	ca	no						
12	0	ca	no						
13	1	ptp	fil						
14	1	ptp	fil						
15	1	ptp	fil						
16	1	ptp	fil						
17	1	my	fil						
18	1	my	fil						
19	1	my	fil						
20	1	my	fil						
21	1	ca	fil						
22	1	ca	fil						
23	1	ca	fil						
24	1	ca	fil						
25	2	ptp	fil						
26	2	ptp	fil						
27	2	ptp	fil						
28	2	ptp	fil						
29	2	my	fil						

data | dictionary | v

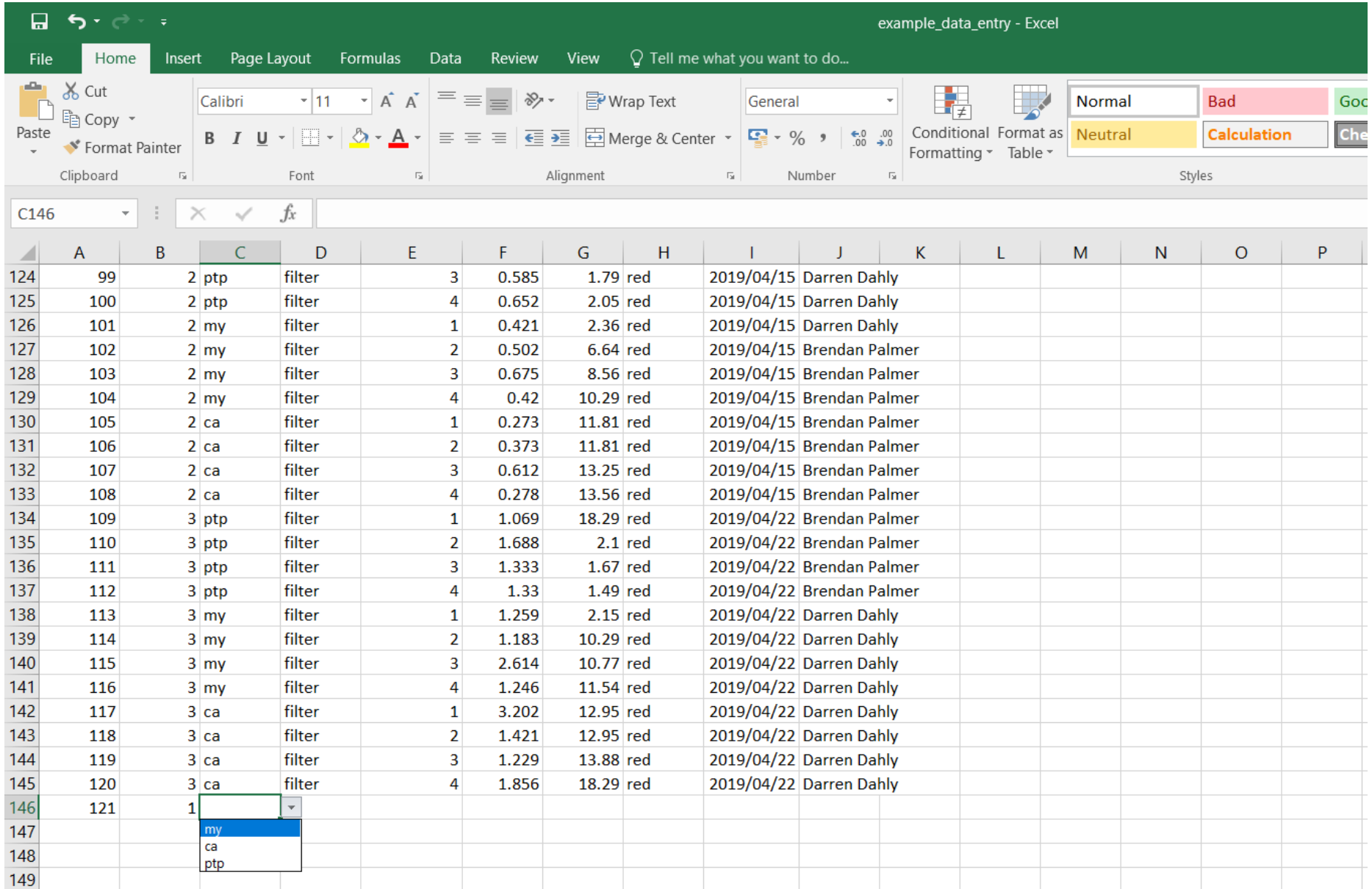
field_name	data_type	data_format	example	standard_units	description
id	numeric	integer	23	NA	Unique identifier applied to each observation
week_no	numeric	integer			
filter_name	character	NA	n		
treatment	character	NA	fi		
replicate_no	numeric	integer			
flavonoids	numeric	double			
biomass	numeric	double			
variety	character	NA	c		
date	date	YYYY/MM/DD			
investigator	character	Firstname Lastname	A		

data | **dictionary** | values

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

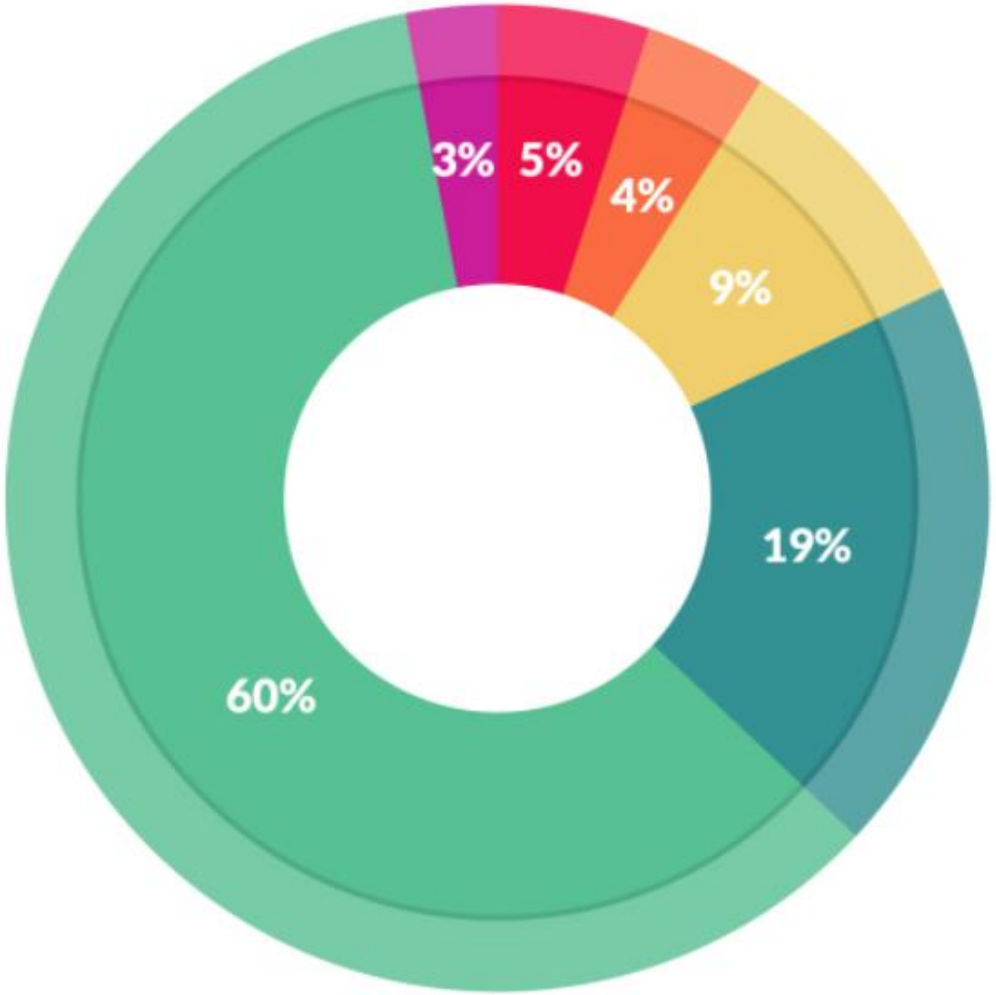
data | dictionary | **values**

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										

Resources are being wasted by not doing this



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

Putting the pieces together

A: Define a project structure

B: Set a naming convention

C: Use scripted workflows

D: Digital notebooks

Reproducible research

E: Version control

F: Data packaging

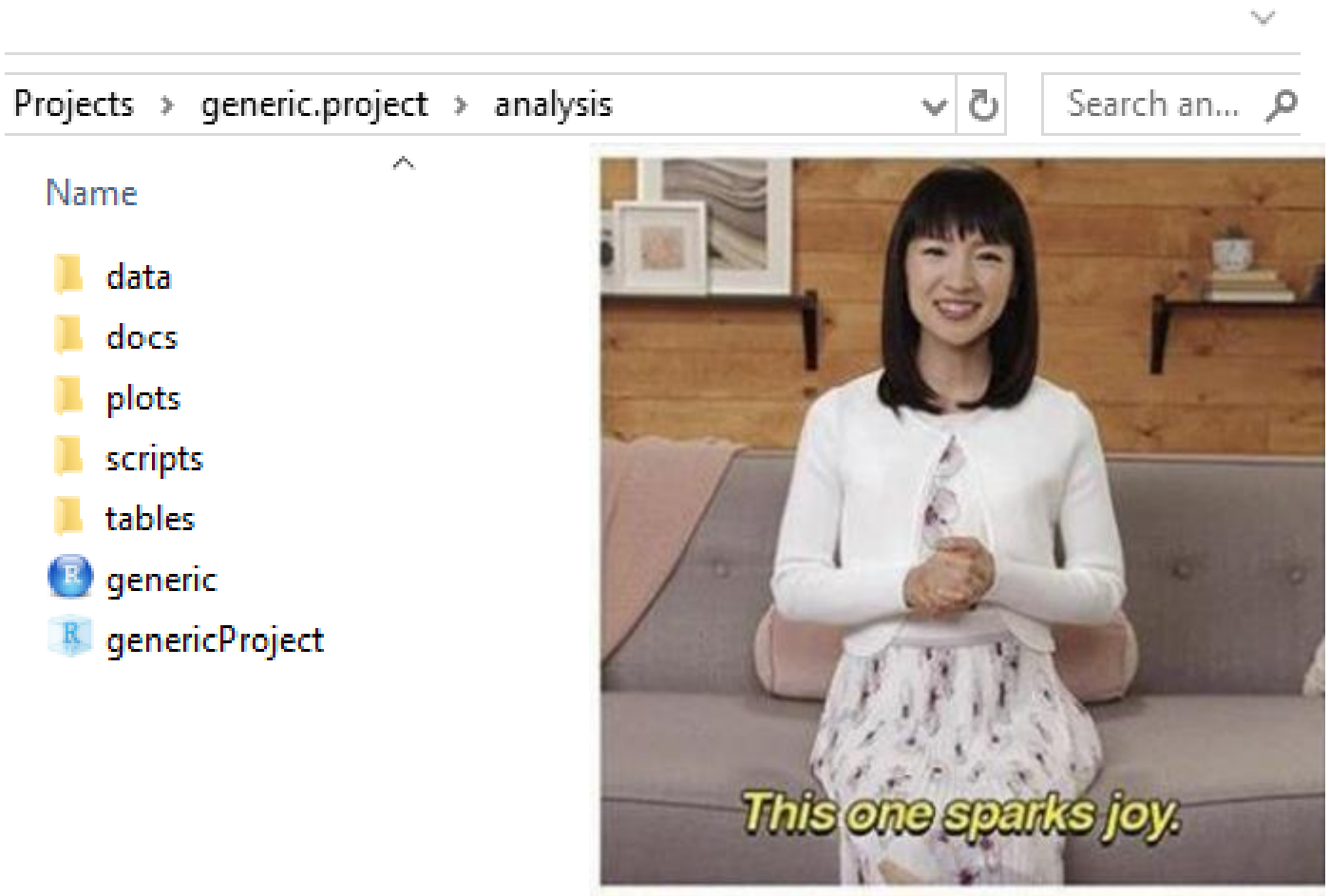
Still haven't found what I'm looking for

- Help your future-self

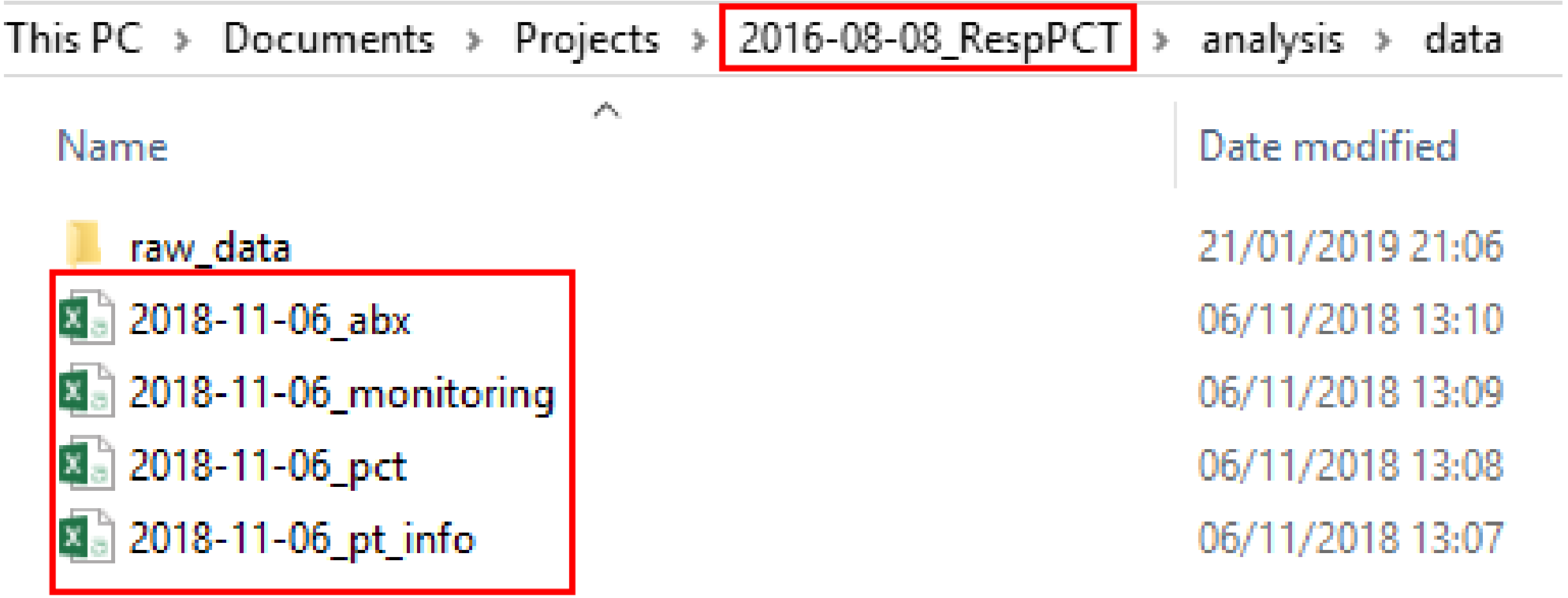
B_Palmer_Medicine_Files › 4a Project › Pyrosequencing_analysis › Pyrosequencing_Paper › Draft_Paper_incl_Figs › Submission › JVI_Resubmission › JVI_resubmission_files › Final Final version

Name	Date modified
Cover_letter_B_A_Palmer_Sept_2014	10/09/2014 17:05
Fig_1_Sept_14	11/09/2014 10:31
Fig_1_Sept_14	10/09/2014 23:07
Fig_2_Sept_14	11/09/2014 10:31
Fig_2_Sept_14	10/09/2014 23:07
Fig_3_Sept_14	11/09/2014 10:31
Fig_3_Sept_14	10/09/2014 23:07
Fig_4_Sept_14	11/09/2014 10:31
Fig_4_Sept_14	10/09/2014 23:07
Fig_5_Sept_14	11/09/2014 10:33
Fig_5_Sept_14	10/09/2014 23:07
HCV_UDPS_B_A_Palmer_Sept_14	17/09/2014 12:21
Response_to_Reviewer_Sept_14	10/09/2014 22:42
Supplementry_Figure_B_A_Palmer_Sept_14	29/08/2014 13:21
Supplementry_Figure_B_A_Palmer_Sept_14	10/09/2014 22:31
Tables_B_A_Palmer_Sept_2014	10/09/2014 22:09

A: Define a generic project structure



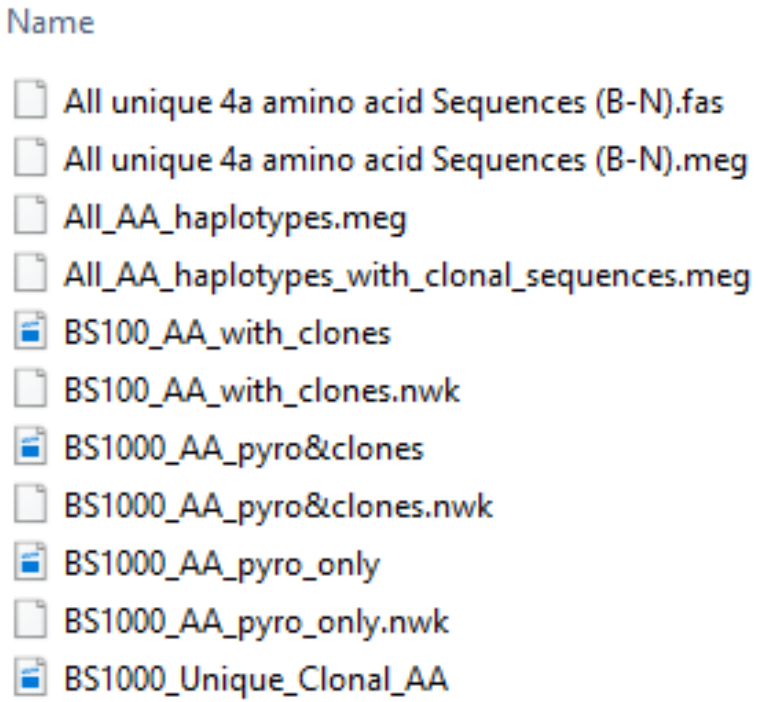
B: Give your files and folders informative names



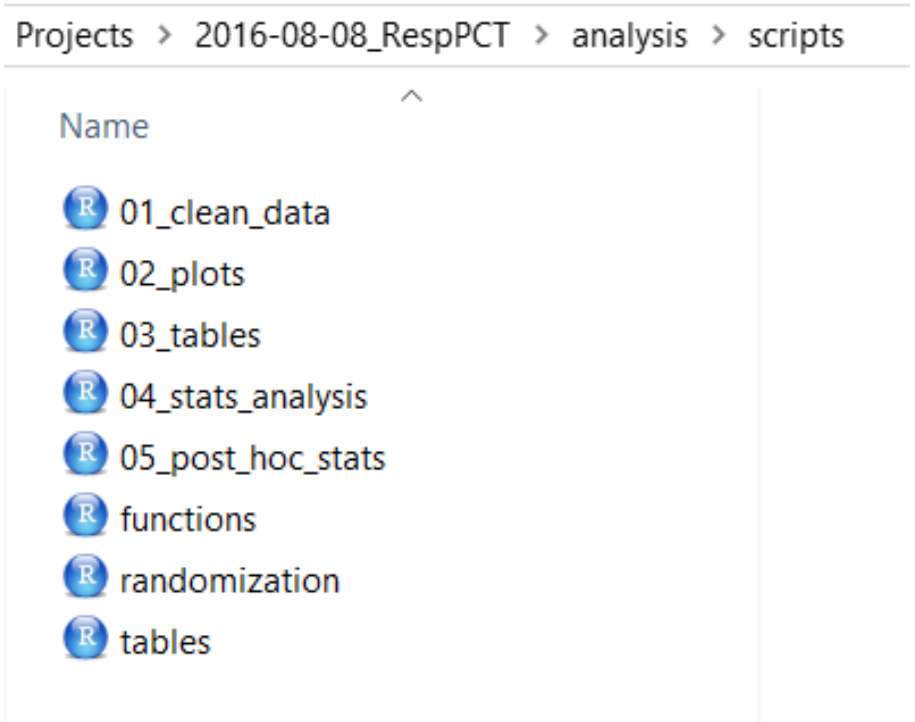
Everything in its right place

- **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

- It can be viewed as an electronic laboratory notebook (ELN)

- 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.

D: R Markdown

~/Open_Science/Digital_Badge/RCR - master - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins

lettuce_report.Rmd*

Knit | Insert | Run

```
---
title: "This is a reproducible document"
author: "Dr. Brendan Palmer"
date: "18th June 2019"
output:
  word_document:
    fig_height: 4
    fig_width: 6
---
# This is the beginning of the project

Our initial reports might be restricted to lab meetings etc. We can use `R Markdown` to show the code we are using, so that the meetings are not just a demonstration of the results, but also an examination of the `code` used to obtain them.

## Data overview
```{r packages and setup, include = FALSE}

knitr::opts_chunk$set(echo = FALSE, message = FALSE, warning = FALSE)

# Load your packages here
library(tidyverse)
library(knitr)
```

The plot below is call from the ggplot object entitled `report_plot` created in the script `03_final_analysis.R`.

```{r Plots from script, echo = FALSE}

source("scripts/03_final_analysis.R")

# The location of the Rmd file dictates whether the path to other files is intact
```

This is a reproducible document

Dr. Brendan Palmer

18th June 2019

This is the beginning of the project

Our initial reports might be restricted to lab meetings etc. We can use R Markdown to show the code we are using, so that the meetings are not just a demonstration of the results, but also an examination of the code used to obtain them.

Data overview

The plot below is call from the ggplot object entitled report_plot created in the script 03_final_analysis.R.

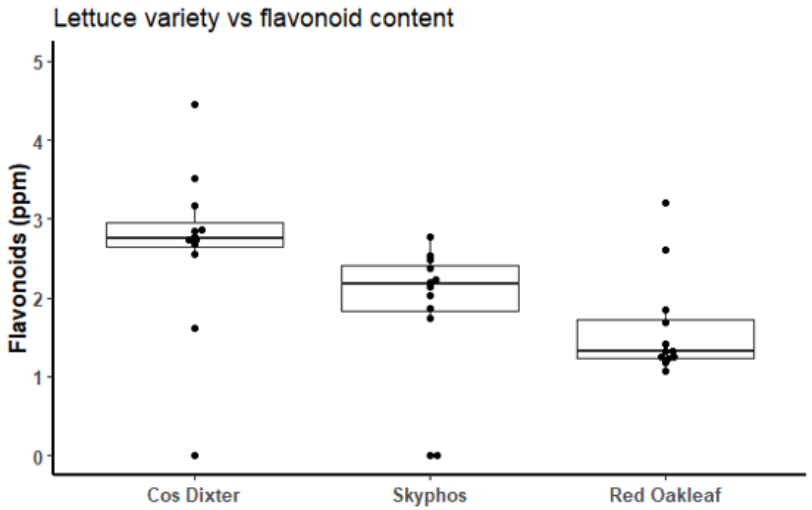
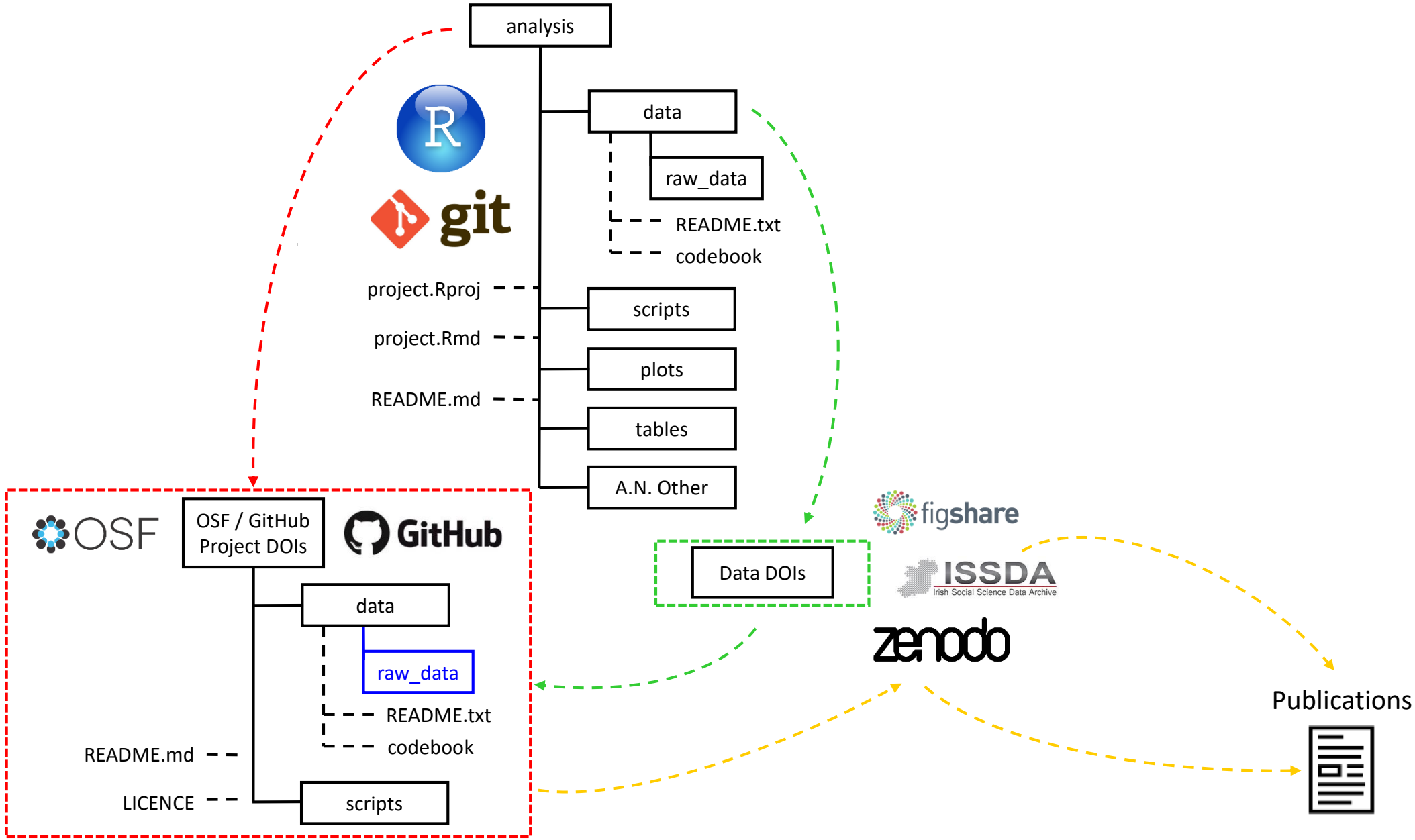


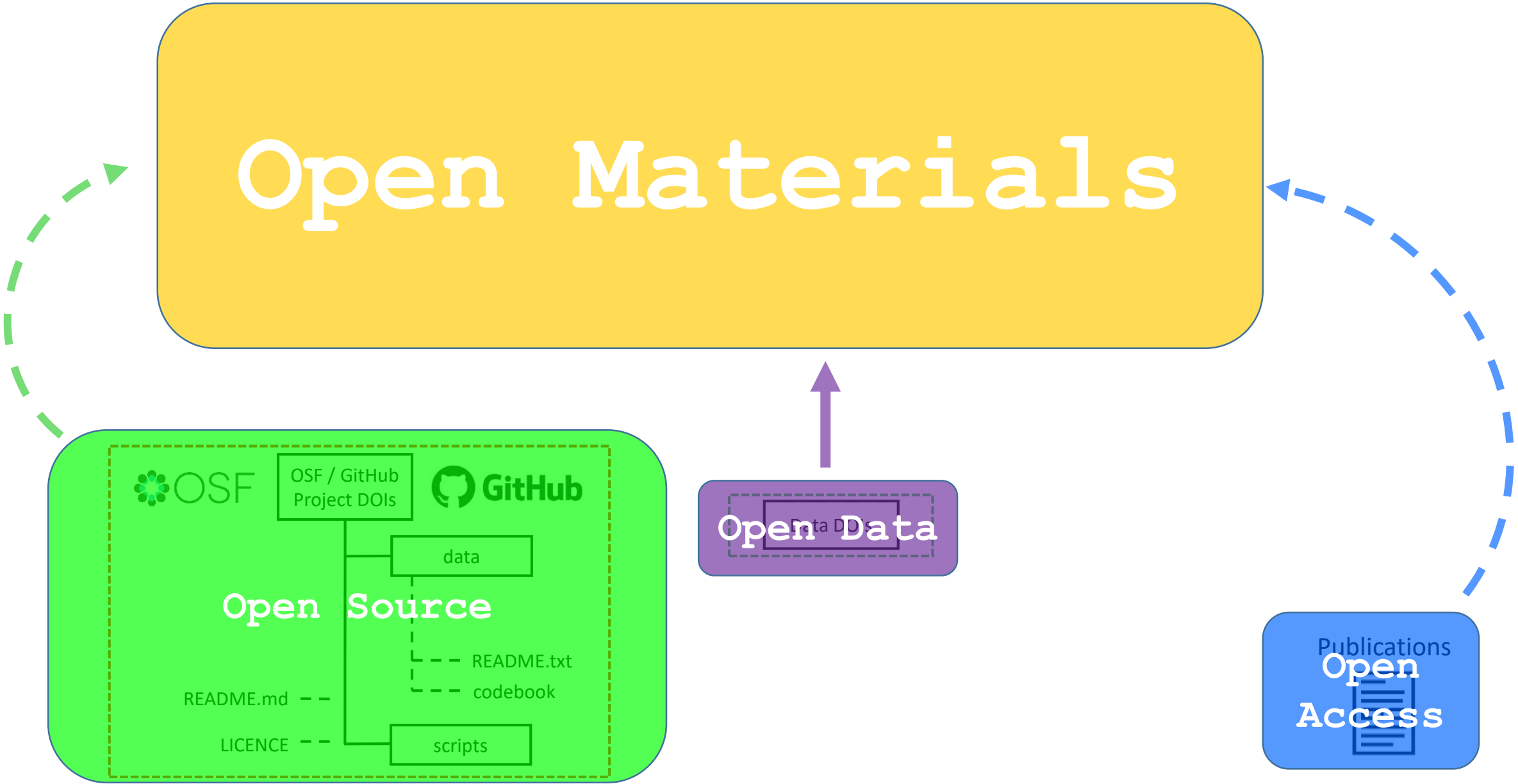
Fig. 1. Flavonoid content of three lettuce varieties under three experimental conditions.

Or we can also recreate the code within the R Markdown document as seen below.
```

What does this allow us to do?



What does this allow us to do?



Mistakes can still happen

Rasmus Nielsen
@ras_nielsen

The one thing that all scientists fear the most is to find out that a major result they have published was based on erroneous data. This is an event that will affect you for the rest of your scientific career. 1/3

6:24 PM · Sep 27, 2019 · Twitter Web App

181 Retweets **1.4K** Likes

Rasmus Nielsen @ras_nielsen · Sep 27
Replying to @ras_nielsen

David Reich (inspired by the work of Sean Harrison) has found an error in the UK Biobank data that likely explains most or all of our results regarding CCR5 delta-32. We will work with the Nature Medicine editors to get the publication record corrected. 2/3

34 95 861

Daniel MacArthur
@dgmacarthur

It's also a good reminder for everyone: never blindly trust any genomic data set. They all contain hidden errors that evade bulk QC, even when very carefully done, but emerge when doing very specific analyses. Be suspicious, and tailor your QC to the question you're asking.

5:31 PM · Oct 2, 2019 · Twitter Web App

Contradictory Results

Comment on this paper

No statistical evidence for an effect of CCR5-Δ32 on lifespan in the UK Biobank cohort

Robert M Maier, Ali Akbari, Xinzhu Wei, Nick Patterson, Rasmus J Nielsen, David E. Reich

doi: https://doi.org/10.1101/787986

This article is a preprint and has not been certified by peer review [what does this mean?].

Abstract | Info/History | Metrics | Preview PDF

Abstract

A recent study reported that a 32-base-pair deletion in the CCR5 gene (CCR5-Δ32) is deleterious in the homozygous state in humans. Evidence for this came from a survival analysis in the UK Biobank cohort, and from deviations from Hardy-Weinberg equilibrium at a polymorphism tagging the deletion (rs62625034). Here, we carry out a joint analysis of whole-genome genotyping data and whole-exome sequencing data from the UK Biobank, which reveals that technical artifacts are a more plausible cause for deviations from Hardy-Weinberg equilibrium at this polymorphism. Specifically, we find that individuals homozygous for the deletion in the sequencing data are underrepresented in the genotyping data due to an elevated rate of missing data at rs62625034, possibly because the probe for this SNP overlaps with the Δ32 deletion. Another variant which has a higher concordance with the deletion in the sequencing data shows no associations with mortality. A phenome-wide scan for effects of variants tagging this deletion shows an overall inflation of association p-values, but identifies only one trait at $p < 5x10-8$, and no mediators for an effect on mortality. These analyses show that the original reports of a recessive deleterious effect of CCR5-Δ32 are affected by a technical artifact, and that a closer investigation of the same data provides no positive evidence for an effect on lifespan.

Subject Area
Genetics

Subject Areas
All Articles
Animal Behavior and Cognition
Biochemistry
Bioengineering
Bioinformatics
Biophysics
Cancer Biology
Cell Biology
Clinical Trials*
Developmental Biology
Ecology
Epidemiology*
Evolutionary Biology
Genetics
Genomics
Immunology
Microbiology

Mistakes can still happen

thebmj

BMJ 2019;367:l6365 doi: 10.1136/bmj.l6365 (Published 21 November 2019)

EDITORIALS

Check for updates

Why researchers should share their analytic code

Retraction of a trial shows the importance of transparency

Ben Goldacre *director*, Caroline E Morton *researcher*, Nicholas J DeVito *researcher*

DataLab, Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK

JAMA recently retracted and replaced an important clinical trial report from 2018 after a serious programming error was discovered.[1] Quantitative medical research relies on analytic scripts: a sequence of commands issued to extract, reshape, manage, and then analyse data. In this case, there was a catastrophe. The "randomisation assignment" variable coded the control group "1" and the intervention group "2"; this had to be converted to "0" and "1" for the statistical analysis to run, but an incorrect conversion command resulted in the intervention and control groups being mislabelled. The results of the trial were almost completely reversed.

the most commonly used, GitHub,[13] has a limit of 100 GB for each repository. For context, our group's OpenPrescribing.net service is a substantial software project with 130 000 users a year: the whole project is over 30 000 lines of code, which is at least one order of magnitude bigger than any single epidemiological analysis script, but this equates to only 1.5 MB of storage.

Another objection is the time needed to create perfectly curated code, but there is no need for code to be converted into generalisable "libraries"; simply sharing practical working code is a good start.[14] Emerging best practice is to share full analyses

The butterfly has started flapping its wings

Plan S
Making full and immediate Open Access a reality

Why Plan S | 10 Principles | Funders & support | Implementation | About | Contact

"After 1 January 2020 scientific publications on the results from research funded by public grants provided by national and European research councils and funding bodies, must be published in compliant Open Access Journals or on compliant Open Access Platforms."

EUROPEAN COMMISSION
Directorate-General for Research & Innovation

H2020 Programme

Guidelines on

FAIR Data Management in Horizon 2020

IRISH RESEARCH COUNCIL
An Chomhairle um Thaighde in Éirinn

f) Data Management Plan – 2 pages max.
- Applicants should address the following issues:
  - What standards will be applied?
  - How will data be exploited and/or shared/made accessible for verification and reuse? If data cannot be made available, why?
  - How data will be curated & preserved?
  - If applicable, how does the applicant plan to make the research data FAIR (findable, accessible, interoperable and reusable).

HRB Health Research Board

Funding | Data collections & evidence | Publications | Success stories | News | About

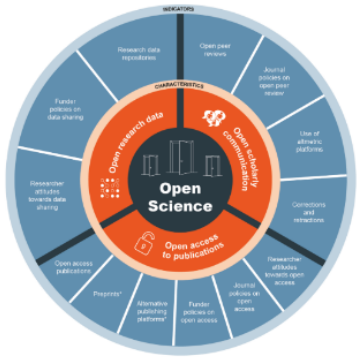
Home > Funding > Policies and principles > Open Research

- Funding schemes
- EU funding support
- Manage a grant
- Funding awarded
- Evaluation
- GDPR guidance for researchers
- Policies and principles
  - EU legislation
  - Gender
  - Good research practice
  - Open Research

Open Research

The HRB is committed to ensuring that its funded research is open, accessible and usable, so it can have the greatest possible impact.

There is a fundamental shift across Europe towards making research more transparent, collaborative, accessible and efficient. This Open Science movement is a strategic priority for the European Commission in research and innovation policy and an EU high-level Expert Group, the Open Science Policy Platform (OSPP 2016-2018) has been established to consider key implementation areas.



Science Foundation Ireland sfi For what's next

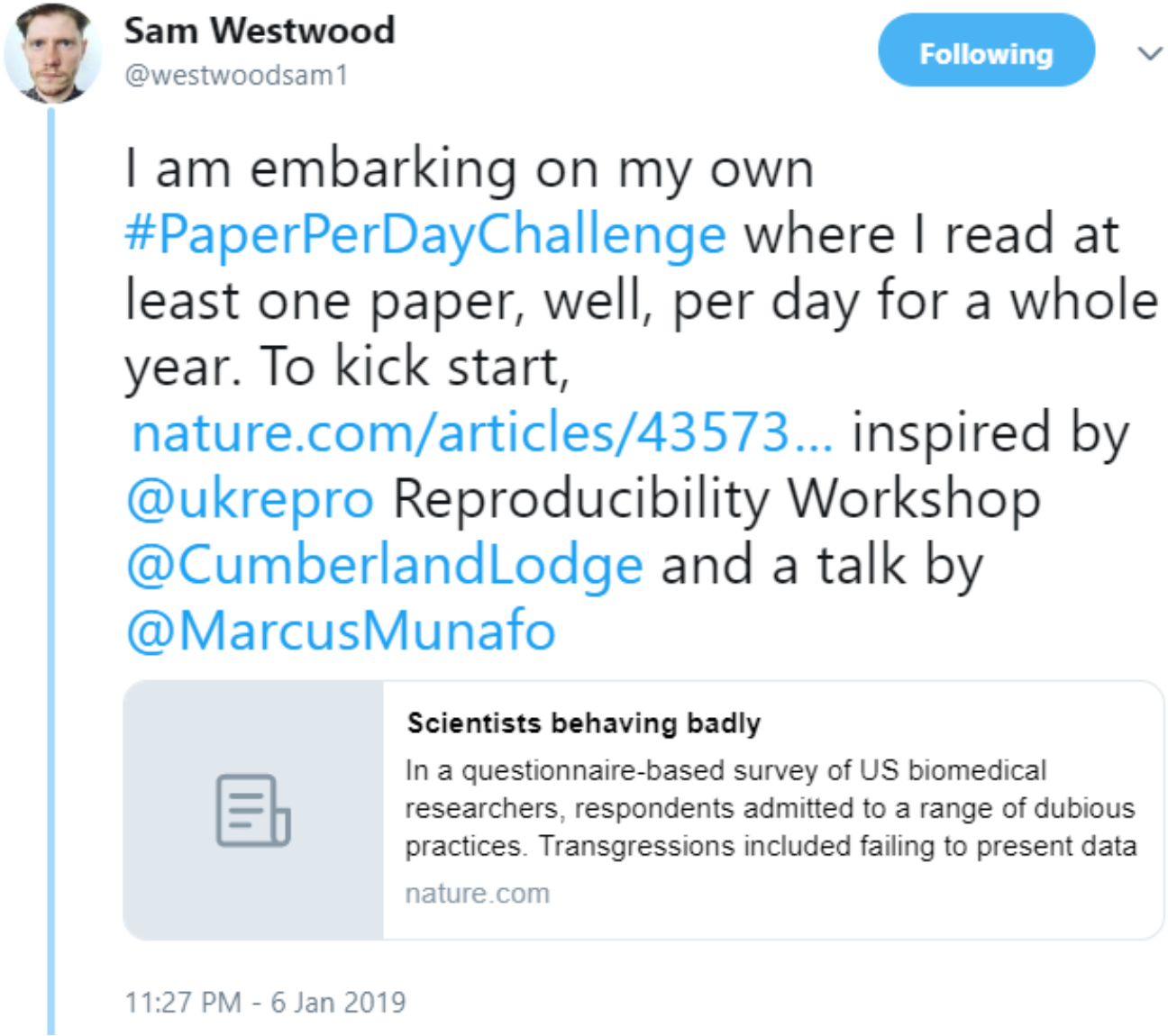
Funding | Engagement | Events | Research News | SFI Research Centres

Science Foundation Ireland joins DORA

14th February 2019, Dublin – Science Foundation Ireland has become a signatory to the San Francisco Declaration of Research Assessment (DORA), making a formal commitment to assessing the quality and impact of research through means other than journal impact factors.

Install the Chrome plugin PubPeer

Further reading



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.

Malcolm Macleod #FBPE
@Maclomaclee Follows you

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

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

Brian Nosek
@BrianNosek

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

Kate Button
@ButtonKate Follows you

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

Darren L Dahly
@statsepi Follows you

Principal Statistician, Epidemiologist, Sr Lecturer | @HRBIreland Clinical Research Facility @CRF_CORK | Cork #Rstats Users Group meetup.com/Cork-Ireland-R...

Dorothy Bishop
@deevybee

Professor of developmental neuropsychology. Blog on deevybee.blogspot.com Main focus #devlangdis, see: youtube.com/radld

Elisabeth Bik
@MicrobiomDigest

Science consultant, PhD. Harbers-Bik LLC. Microbiome, research integrity & misconduct. Ex @Stanford. MicrobiomeDigest/Bik's Picks. Dutch/USA. My views.

Retraction Watch
@RetractionWatch

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

Jenny Bryan
@JennyBryan

Software engineer @rstudio, humane #rstats, adjunct prof @UBC where I created @STAT545, part of @ropensci

Coming to a UCC near you in 2020

Brendan Palmer
@B_A_Palmer

🚨 BIG NEWS: We've just received approval for a new postgraduate module where we'll teach reproducible scripted workflows through #rstats!

Thanks to all who helped get it to this point, but special mention to @statsepi who set the ball rolling 🏝️

@UCC > Modules > PG6030 🥳

Module Content: The module will introduce fundamental concepts of reproducible research alongside hands-on training in the R programming language. Students will be instructed on data collation, curation and management techniques that will serve as a foundation towards downstream visualisation, analysis and reporting via scripted, reproducible workflows.