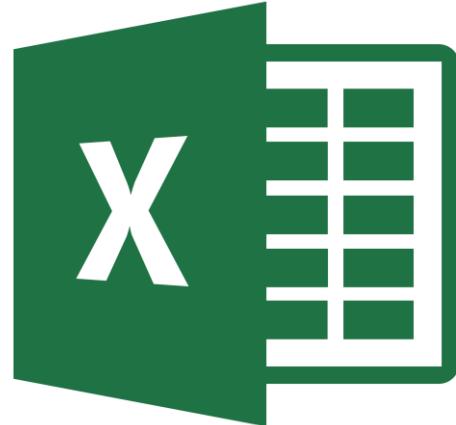


Some of my best friends  
use spreadsheets



---

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

 @B\_A\_Palmer



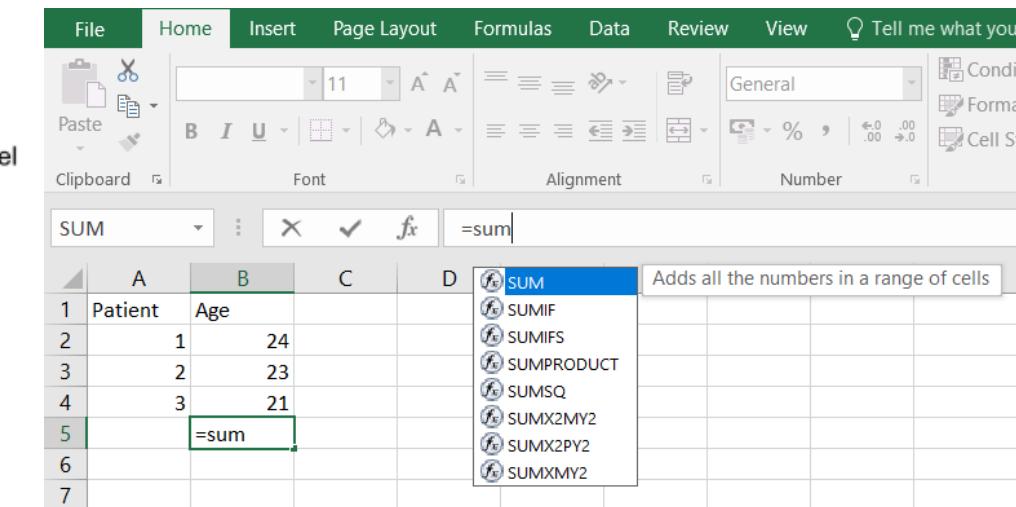
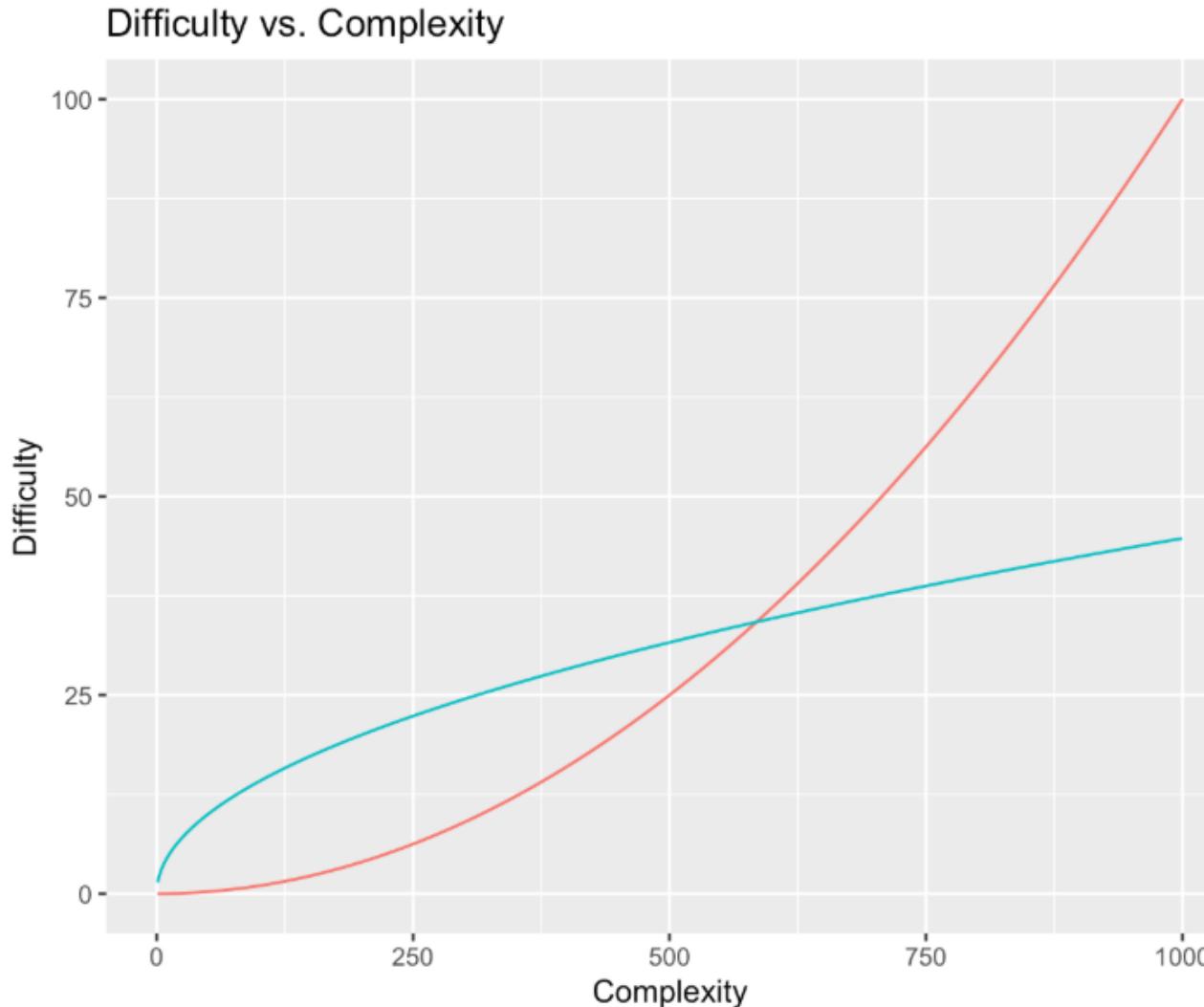
**Andrew Althouse**  
@ADAlthousePhD



Statistician's tip of the day: if someone tells you that they "just have a few Excel sheets" that they want help with, run the other way. Also, you may want to give them a fake phone number, possibly a fake name. It may be worth faking your own death, in extreme circumstances.

3:39 PM · Jul 11, 2018 · [Twitter Web Client](#)

# Excel is intuitive to use



# But a breeding ground for errors

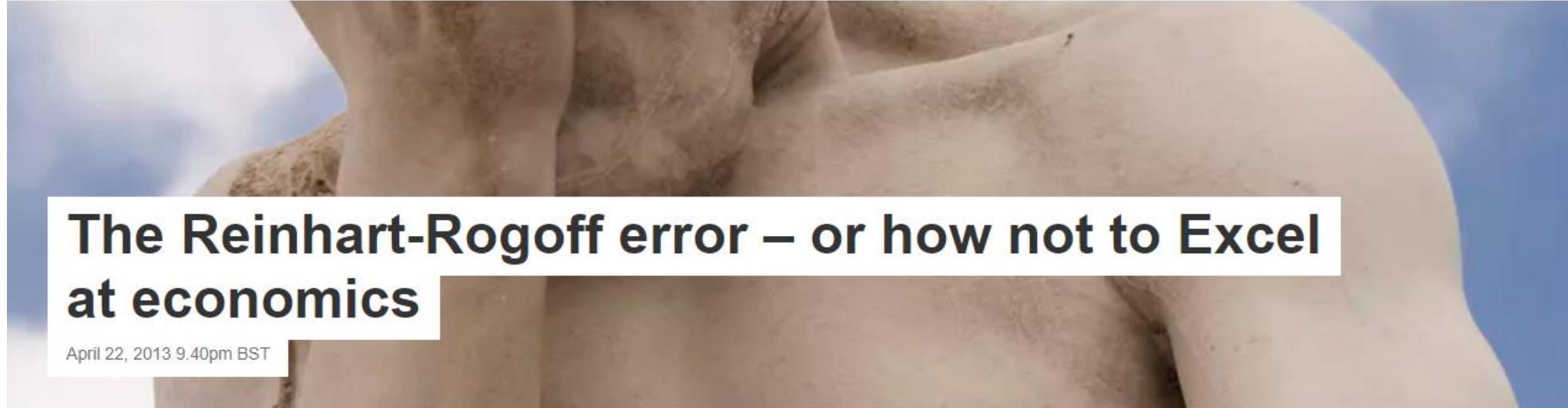
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Patient	Age	Group	t.test									
2	1	24	A										
3	2	23	A										
4	3	21	A										
5	4	45	B										
6	5	36	B										
7	6	68	B	=T.TEST(B2:B4,B5:B7, 1)									
8				T.TEST(array1, array2, tails, type)	red								
9													
10													

	A	B	C	D	E
1	Patient	Age	Group	t.test	
2	1	24	A		
3	2	23	A		
4	3	21	A		
5	4	45	B		
6	5	36	B		
7	6	68	B	0.0596	✓✓✓✓✓
8				0.023871	✓✓✓✓✓
9				0.051999	
10					

	A	B	C	D	E
1	Patient	Age	Group	t.test	
2	1	24	A		
3	2	23	A		
4	3	21	A		
5	4	45	B		
6	5	36	B		
7	6	68	B	=T.TEST(B2:B4,B5:B6, 1, 1)	
8				T.TEST(array1, array2, tails, type)	red
9					
10					

	A	B	C	D	E
1	Patient	Age	Group	t.test	
2	1	24	A		
3	2	23	A		
4	3	21	A		
5	4	45	B		
6	5	36	B		
7	6	68	B	#N/A	
8				0.007552	✓✓✓✓✓
9				0.073071	
10					

# And mistakes can happen all too easily



## 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](#)

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.

# This is a big worldwide problem

# Fundamental problem – default settings



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



CrossMark

## Gene name errors are widespread in the scientific literature

Mark Ziemann<sup>1</sup>, Yotam Eren<sup>1,2</sup> and Assam El-Osta<sup>1,3\*</sup>

### 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-

- Why did it take us until 2016 to discover this?

\*cough\* We've known for a long time \*cough\*

**BMC Bioinformatics**



Correspondence

Open Access

**Mistaken Identifiers: Gene name errors can be introduced inadvertently when using Excel in bioinformatics**

Barry R Zeeberg<sup>†1</sup>, Joseph Riss<sup>†2</sup>, David W Kane<sup>3</sup>, Kimberly J Bussey<sup>1</sup>, Edward Uchio<sup>4</sup>, W Marston Linehan<sup>4</sup>, J Carl Barrett<sup>2</sup> and John N Weinstein<sup>\*1</sup>

2004

Figure 1

	gene names	internal date format	default date format	gene names	internal date format	default date format	gene names	internal date format	default date format
1	APR-1	35885	1-Apr	OCT-1	36068	1-Oct	SEP2	36039	2-Sep
2	APR-2	35886	2-Apr	OCT-2	36069	2-Oct	SEP3	36040	3-Sep
3	APR-3	35887	3-Apr	OCT-3	36070	3-Oct	SEP4	36041	4-Sep
4	APR-4	35888	4-Apr	OCT-4	36071	4-Oct	SEP5	36042	5-Sep
5	APR-5	35889	5-Apr	OCT-6	36073	6-Oct	SEP6	36043	6-Sep
6	DEC-1	36129	1-Dec	OCT1	36068	1-Oct	SEPT1	36038	1-Sep
7	DEC-2	36130	2-Dec	OCT11	36078	11-Oct	SEPT2	36039	2-Sep
8	DEC1	36129	1-Dec	OCT2	36069	2-Oct	SEPT3	36040	3-Sep
9	DEC2	36130	2-Dec	OCT3	36070	3-Oct	SEPT4	36041	4-Sep
10	MAR1	35854	1-Mar	OCT4	36071	4-Oct	SEPT5	36042	5-Sep
11	MAR2	35855	2-Mar	OCT6	36073	6-Oct	SEPT6	36043	6-Sep
12	MAR3	35856	3-Mar	OCT7	36074	7-Oct	SEPT7	36044	7-Sep
13	NOV1	36099	1-Nov	SEP-1	36038	1-Sep	SEPT8	36045	8-Sep
14	NOV2	36100	2-Nov	SEP-2	36039	2-Sep	SEPT9	36046	9-Sep
15				SEP1	36038	1-Sep			

# But it doesn't end there

Date and time expressed according to ISO 8601 [\[refresh\]](#)

Date	2019-10-15
Date and time in UTC	2019-10-15T19:49:52+00:00
Week	2019-W42
Date with week number	2019-W42-2
Date without year	--10-15 <sup>[1]</sup>
Ordinal date	2019-288

- YYYY-MM-DD or YYYYMMDD

- Type this into Excel

A	B
1	2019-10-15
2	

- And hit return

A	B
1	15/10/2019
2	

- DD/MM/YYYY

# Excel also frequently gets clipboard amnesia



The answer, unfortunately, is **no**, you can't stop this from happening.

25

As described by [Joel Spolsky](#), developer and program manager for excel:



The official reason is that Excel doesn't really have cut and paste, it has move and copy. That's necessary because Excel automatically does reference fix up. For example, if cell A2 is defined as =A1, and you move cell A1 to A3, cell A2 will be updated to =A3.



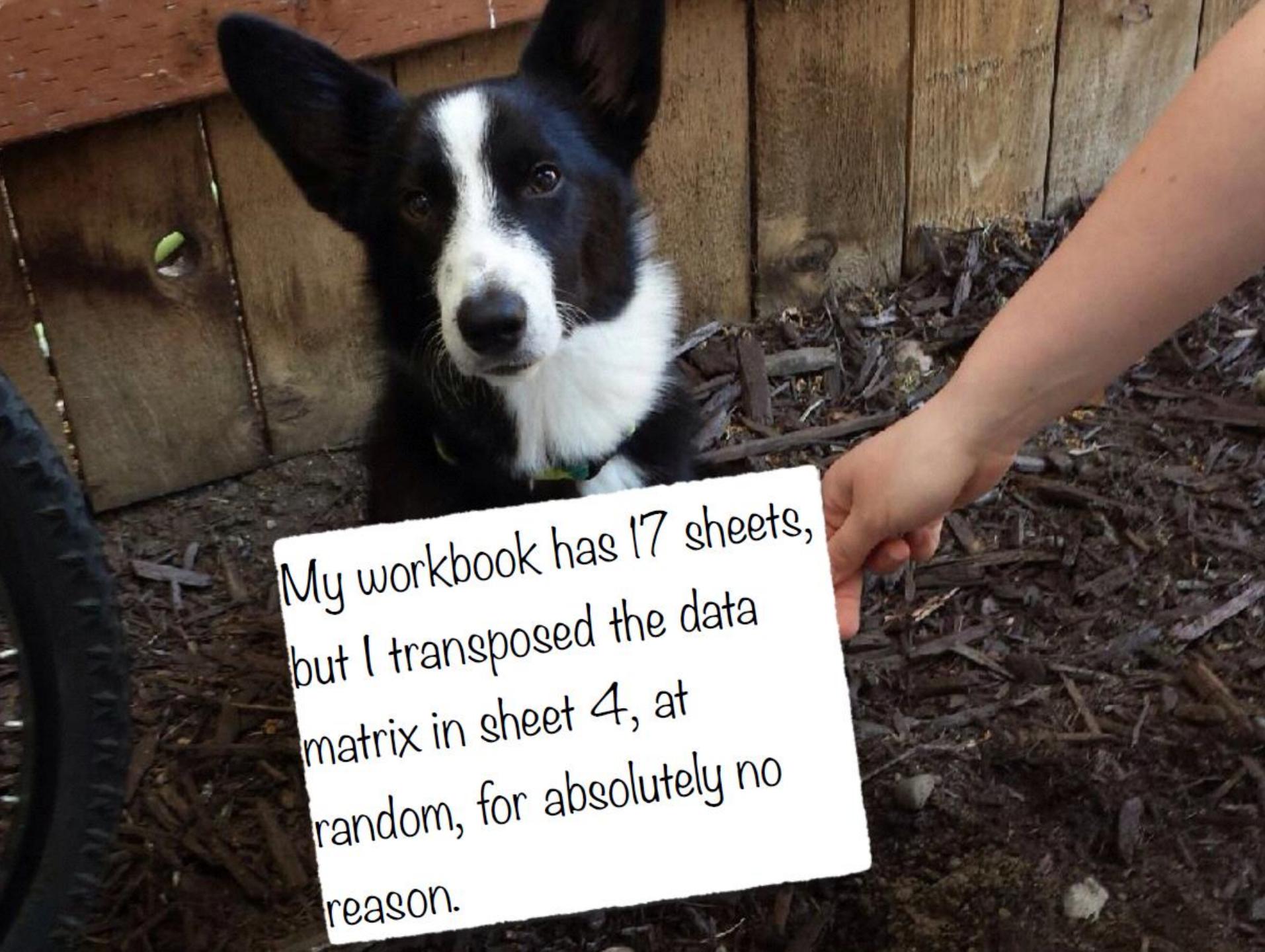
If Excel actually cut things to the clipboard you would somehow need to have a reference pointing >into< the clipboard which is bizarre and for which there is no reasonable syntax. In other words, Excel doesn't want to leave you with dangling references during a move operation and isn't confident that it would be able to fix them up correctly when you completed the move by selecting "Paste."

Joel Spolsky 3/9/2004

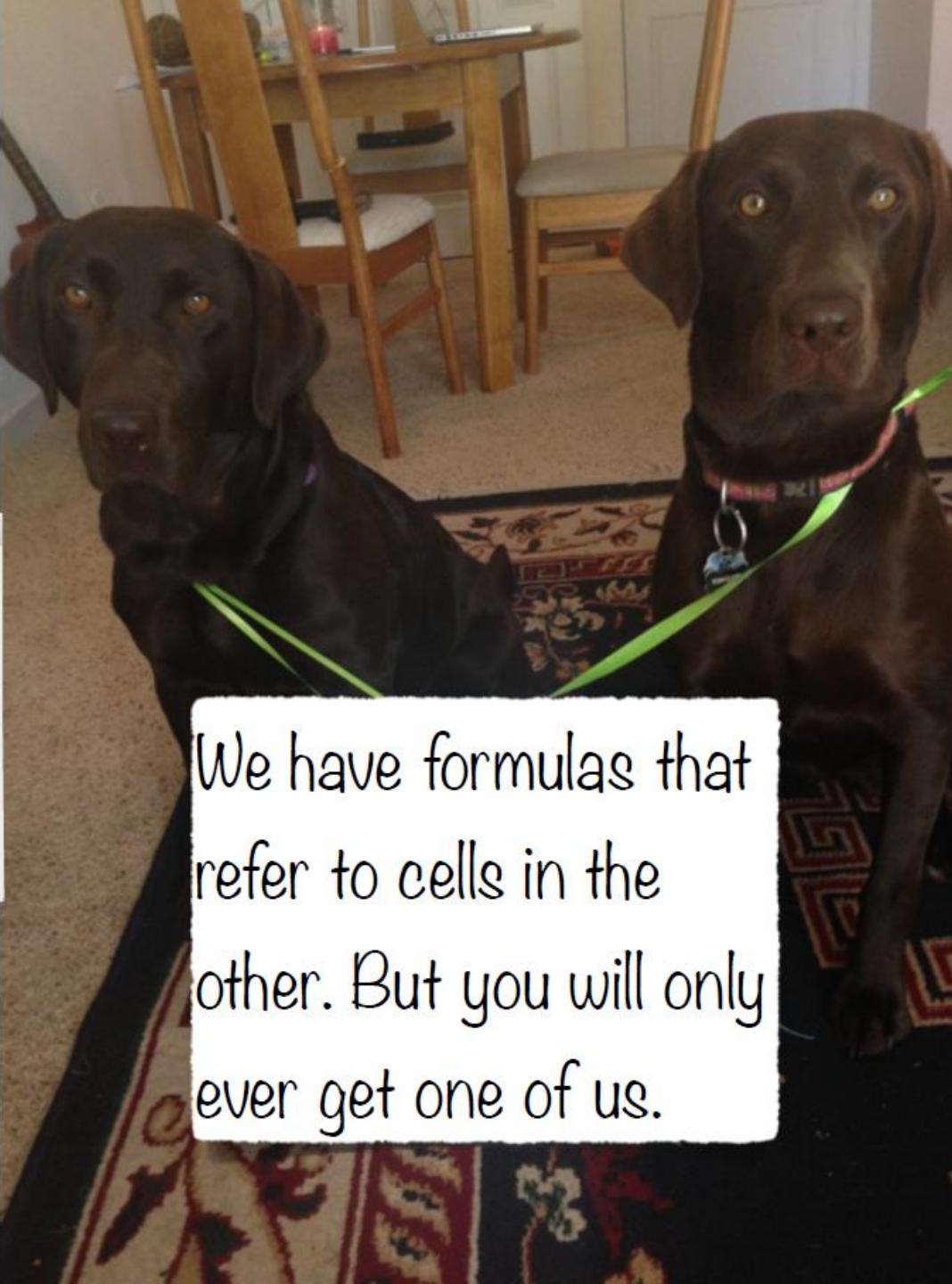
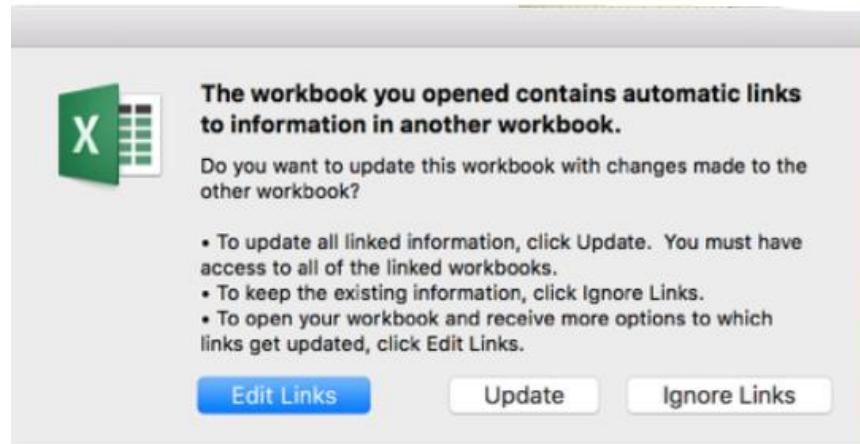
[source](#)

What this means is that because of the difficulty inherent in the way excel maintains *references*, at the time of development there was no good way to store these references outside of excel and have them remain dynamic to be re-inserted. Once you change *focus* excel's ability to retain your original references is lost.

Unfortunately, MS does not consider this a bug.

A black and white dog with a white patch on its chest is looking towards a person's hand. The person is holding a white piece of paper with handwritten text on it. The background shows a wooden fence and some mulch or straw on the ground.

My workbook has 17 sheets,  
but I transposed the data  
matrix in sheet 4, at  
random, for absolutely no  
reason.





Columns of intermediate computations are so boring. I like to hide them!

File Home Insert Page Layout Formulas Data Review View Design Format

Tell me what you want to do...

Cut Copy Paste Format Painter

Font Alignment Number Styles Cells

AutoSum Fill Sort & Filter Clear

Clipboard

A B C D E F G H I J K L M N O P Q R S

22

23 Amino Acid Complexity Across Fractions

24 Species Name Fraction Sequence Length (bp) Original Clone

25 HUF-1 UF 318 HUF-1/HUF-2/HUF-3/HUF-4/HUF-5/HUF-6/HUF-7/HUF-8/HUF-9/HUF-10

26 HUF-2 UF 321 HUF-13/HUF-15/HUF-16/HUF-17/HUF-18/HUF-19/HUF-20/HUF-21/HUF-22/HUF-23/HUF-24/HUF-7/HUF-11/HUF-18/HUF-10/HUF-16/HUF-17/HUF-23/HUF-24/HUF-1/HUF-3/HUF-6/HUF-8/HUF-12/HUF-14

27 HU-1 U1 318 HU-1/HU-3/HU-4/HU-10/HU-12/HU-13/HU-14/HU-17/HU-18/HU-22

28 HU-2 U1 321 HU-2/HU-5/HU-6/HU-7/HU-11/HU-19/HU-22/HU-24

29 HU-3 U1 321 HU-20

30 HU-4 U1 321 HU-8

31 HB-1 B 321 HB-2/HB-4/HB-5/HB-6/H-8/HB-10

32 HB-2 B 321 HB-3/HB-7/HB-9

33

34

35 Unfractionated IgG-depleted IgG-enriched Colour

36 HUF-1 10

37 HUF-2 32

38 HU-1 10

39 HU-2 13

40 HU-3 1

41 HU-4 1

42 HB-1 6

43 HB-2 3

44

45

46

47

48

49

50

51

52

H Complexity at AA Level

Fractionation Type	Unfractionated	IgG-depleted	IgG-enriched
Unfractionated	~22%	~22%	~56%
IgG-depleted	~40%	~50%	~10%
IgG-enriched	~65%	~30%	~5%

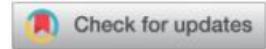
B 4395 | E 6576 | O 6998 | F 7386 | H 9574 | I 11115 | J 11674 | L 12888 | M 14025 | QS evolution at nuc minus ... | + | Ready

# Take small steps to big changes

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



OPEN ACCESS



## Data Organization in Spreadsheets

Karl W. Broman<sup>a</sup> and Kara H. Woo<sup>b</sup>

<sup>a</sup>Department of Biostatistics & Medical Informatics, University of Wisconsin-Madison, Madison, WI; <sup>b</sup>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

# 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

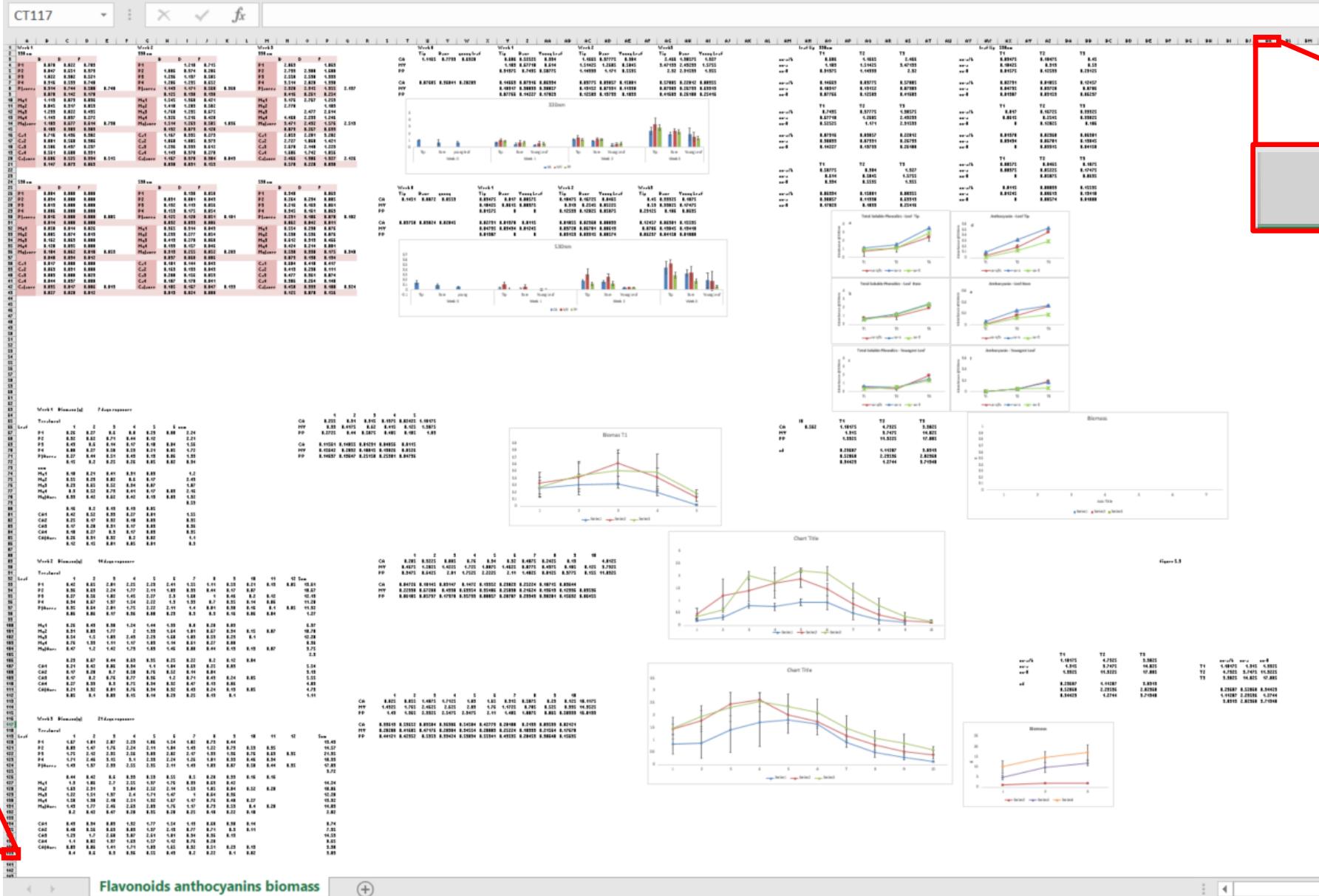
Raw Data wk 1-3 Lettuce Exp 1 - Excel

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Week 1						Week 2						Week 3					
2	330 nm						330 nm						330 nm					
3		B	D	F				B	D	F				B	D	F		
4	P1	0.870	0.822	0.703			P1						1	2.869		1.069		
5	P2	0.847	0.651	0.379			P2						2	2.739	2.380	1.688		
6	P3	1.022	0.902	0.521			P3	1.236	1.197	0.585			P3	2.558	2.538	1.333		
7	P4	0.916	0.599	0.748			P4	1.206	1.295	0.652			P4	3.514	2.028	1.330		
8	P(average)	0.914	0.744	0.588	0.748		P(average)	1.149	1.171	0.560	0.960		P(average)	2.920	2.315	1.355	2.197	
9		0.078	0.142	0.170				0.125	0.138	0.190				0.416	0.261	0.254		
10	My1	1.119	0.873	0.896			My1	1.545	1.360	0.421			My1	3.176	2.767	1.259		
11	My2	0.845	0.917	0.853			My2	1.418	1.203	0.502			My2	2.778		1.183		
12	My3	1.299	0.822	0.435			My3	1.768	1.295	0.675			My3		2.477	2.614		
13	My4	1.149	0.097	0.272			My4	1.326	1.216	0.420			My4	4.460	2.233	1.246		
14	My(average)	1.103	0.677	0.614	0.798		My(average)	1.514	1.269	0.505	1.096		My(average)	3.471	2.492	1.576	2.513	
15		0.189	0.389	0.309				0.192	0.073	0.120				0.879	0.267	0.693		
16	Ca1	0.716	0.496	0.382			Ca1	1.167	0.935	0.273			Ca1	2.853	2.201	3.202		
17	Ca2	0.881	0.568	0.386			Ca2	1.060	1.005	0.373			Ca2	2.727	1.860	1.421		
18	Ca3	0.586	0.437	0.237			Ca3	1.296	0.993	0.612			Ca3	2.678	2.140	1.229		
19	Ca4	0.561	0.600	0.331			Ca4	1.143	0.978	0.278			Ca4	1.606	1.742	1.856		
20	Ca(average)	0.686	0.525	0.334	0.515		Ca(average)	1.167	0.978	0.384	0.843		Ca(average)	2.466	1.986	1.927	2.126	
21		0.147	0.073	0.069				0.098	0.031	0.159				0.578	0.220	0.890		
22																		
23																		
24	530 nm						530 nm						530 nm					
25		B	D	F				B	D	F				B	D	F		
26	P1	0.004	0.000	0.000			P1		0.138	0.050				P1	0.340		0.069	
27	P2	0.034	0.000	0.000			P2		0.091	0.081	0.043			P2	0.264	0.234	0.085	CA
28	P3	0.019	0.000	0.000			P3		0.132	0.119	0.056			P3	0.216	0.163	0.061	MY

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

Normal Page Break Preview Custom Layout Views Workbook Views

Ruler Formula Bar Gridlines Headings Zoom 100% Zoom to Selection Window New Arrange Freeze All Panes Hide Synchronous Scrolling Reset Window Position Window Switch Windows Macros Macros



Raw Data wk 1-3 Lettuce Exp 1 - Excel

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 Layout Custom Views Workbook Views

Ruler Formula Bar Gridlines Headings

Zoom 100% Zoom to Selection Window All Freeze Panes Unhide Synchronous Scrolling Switch Windows Macros

CT117

Figure 5.3 displays a Microsoft Excel spreadsheet titled "Raw Data wk 1-3 Lettuce Exp 1 - Excel". The spreadsheet contains four main sections: "Flavonoids anthocyanins biomass", "Total soluble phenolics", "Biomass", and "Antioxidant activity".

The "Flavonoids anthocyanins biomass" section includes tables for Week 1, Week 2, and Week 3, showing biomass values for various treatments (T1, T2, T3, T4) across different weeks (1-12). It also features four line graphs showing the trend of flavonoid and anthocyanin biomass over time for each treatment.

The "Total soluble phenolics" section includes tables for Week 1, Week 2, and Week 3, showing phenolic content for various treatments (T1, T2, T3, T4) across different weeks (1-12). It also features four line graphs showing the trend of total soluble phenolics over time for each treatment.

The "Biomass" section includes tables for Week 1, Week 2, and Week 3, showing biomass values for various treatments (T1, T2, T3, T4) across different weeks (1-12). It also features four line graphs showing the trend of biomass over time for each treatment.

The "Antioxidant activity" section includes tables for Week 1, Week 2, and Week 3, showing antioxidant activity for various treatments (T1, T2, T3, T4) across different weeks (1-12). It also features four line graphs showing the trend of antioxidant activity over time for each treatment.

## rna\_data - Excel

	A	B	C	D	E
1	GID	YORF	NAME	GWEIGHT	GO.05
2	GENE1331X	A_06_P5820	SFB2    ER to Golgi transport    molecular function unknown    YNL049C    1082129	1	-0.24
3	GENE4924X	A_06_P5866	biological process unknown    molecular function unknown    YNL095C    1086222	1	0.28
4	GENE4690X	A_06_P1834	QRI7    proteolysis and peptidolysis    metalloendopeptidase activity    YDL104C    1085955	1	-0.02
5	GENE1177X	A_06_P4928	CFT2    mRNA polyadenylation*    RNA binding    YLR115W    1081958	1	-0.33
6	GENE511X	A_06_P5620	SSO2    vesicle fusion*    t-SNARE activity    YMR183C    1081214	1	0.05
7	GENE2133X	A_06_P5307	PSP2    biological process unknown    molecular function unknown    YML017W    1083036	1	-0.69
8	GENE1002X	A_06_P6258	RIB2    riboflavin biosynthesis    pseudouridylate synthase activity*    YOL066C    1081766	1	-0.55
9	GENE5478X	A_06_P7082	VMA13    vacuolar acidification    hydrogen-transporting ATPase activity, rotational mechanism    YPR036W    10	1	-0.75
10	GENE2065X	A_06_P2554	EDC3    deadenylylation-independent decapping    molecular function unknown    YEL015W    1082963	1	-0.24
11	GENE2440X	A_06_P6431	VPS5    protein retention in Golgi*    protein transporter activity    YOR069W    1083389	1	-0.16

# Tidy data is clean data



---

*Journal of Statistical Software*

MMMMMM YYYY, Volume VV, Issue II.

<http://www.jstatsoft.org/>

---

## Tidy Data

Hadley Wickham  
RStudio

- Each variable forms a column
- Each observation forms a row
- Each cell contains a value

# 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	mv	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_name	treatment	replicate_no	flavonoids	biomass	variety	date	investigator	
2	1	0	ptp	no	1	1.051	0.30	---	2019/06/28	Aoife Coffey	
3	2	0	ptp	no	A	B	C	D	E		
4	3	0	ptp	no	1	field_name	data_type	data_format	example	standard_units	description
5	4	0	ptp	no	2	id	numeric	integer	23	NA	Unique identifier applied to each observation
6	5	0	my	no	3	week_no	numeric	integer	1	NA	Week number, 1 = 7 days exposure, 2 = 14 days exposure
7	6	0	my	no	4	filter_name	character	NA	my	NA	3 filter types; 'ptp' = polytunnel plastic blocks all UV light
8	7	0	my	no	5	treatment	character	NA	filter	NA	Presence or absence of a filter at the time of sampling
9	8	0	my	no	6	replicate_no	numeric	integer	1	NA	The number of replicates in each treatment
10	9	0	ca	no	7	flavonoids	numeric	double	0.3421	parts per million (ppm)	Leaf disc taken from the tip of the most mature leaf at th
11	10	0	ca	no	8	biomass	numeric	double		gram (g)	Above ground biomass on the day of harvest
12	11	0	ca	no	9	variety	character	NA	cos	NA	3 commerical varieties of red lettuce used; 'cos' = Cos Di
13	12	0	ca	no	10	date	date	YYYY/MM/DD	2019/06/28	ISO 8601	Experiment date
14	13	1	ptp	fil	11	investigator	character	Firstname Lastname	Aoife Coffey	NA	Primary researcher who performed the experiment
15	14	1	ptp	fil	12						
16	15	1	ptp	fil	13						
17	16	1	ptp	fil	14						
18	17	1	my	fil	15						
19	18	1	my	fil	16						
20	19	1	my	fil	17						
21	20	1	my	fil	18						
22	21	1	ca	fil	19						
23	22	1	ca	fil	20						
24	23	1	ca	fil	21						
25	24	1	ca	fil	22						
26	25	2	ptp	fil	23						
27	26	2	ptp	fil	24						
28	27	2	ptp	fil	25						
29	28	2	ptp	fil	26						
30	29	2	mv	fil	27						
					28						
					29						
					30						

# Less stress, more success

The screenshot shows the Quaise software interface with two main tables displayed side-by-side.

**Left Table (data):**

	A	B	C	D	E	F	G	H	I	J	K
1	id	week_no	filter_name	treatment	replicate_no	flavonoids	biomass	variety	date	investigator	
2	1	0	ptp	no	1	1.051	0.20	cos	2010/01/01	Brendan Palmer	
3	2	0	ptp	no	A						
4	3	0	ptp	no	B						
5	4	0	ptp	no	C						
6	5	0	my	no	1	field_name	data_type	data_format	example	standard_units	description
7	6	0	my	no	2	id	numeric	integer			
8	7	0	my	no	3	week_no	numeric	integer			
9	8	0	my	no	4	filter_name	character	NA			
10	9	0	ca	no	5	treatment	character	NA			
11	10	0	ca	no	6	replicate_no	numeric	integer			
12	11	0	ca	no	7	flavonoids	numeric	double			
13	12	0	ca	no	8	biomass	numeric	double			
14	13	1	ptp	fil	9	variety	character	NA			
15	14	1	ptp	fil	10	date	date	YYYY/MM/DD			
16	15	1	ptp	fil	11	investigator	character	Firstname Lastname	A		
17	16	1	ptp	fil	12						
18	17	1	my	fil	13						
19	18	1	my	fil	14						
20	19	1	my	fil	15						
21	20	1	my	fil	16						
22	21	1	ca	fil	17						
23	22	1	ca	fil	18						
24	23	1	ca	fil	19						
25	24	1	ca	fil	20						
26	25	2	ptp	fil	21						
27	26	2	ptp	fil	22						
28	27	2	ptp	fil	23						
29	28	2	ptp	fil	24						
30	29	2	mv	fil	25						
				data	26						
				dictionary	27						
				v	28						
					29						
					30						

**Bottom Navigation:** data | dictionary | values | +

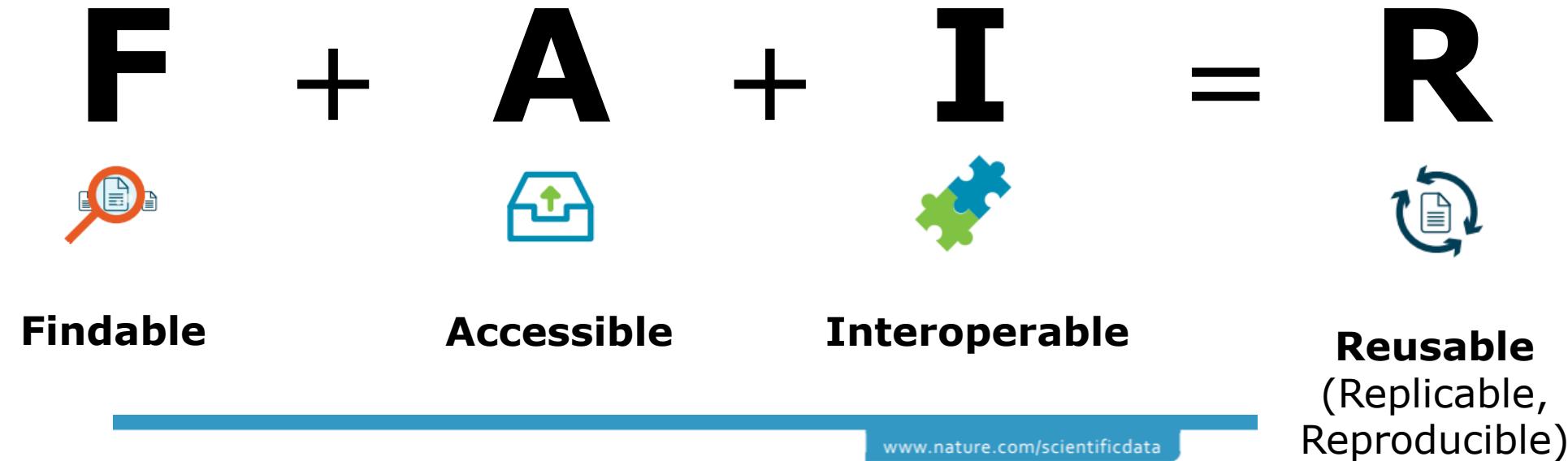
**Right Table (values):**

	A	B	C	D	E	F	G	H	I	J	K
1	id	week_no	filter_name	treatment	replicate_no	flavonoids	biomass	variety	date	investigator	
2			0 my	filter	1						
3			1 ca	no_filter	2			cos		Brendan Palmer	
4			2 ptp		3			sky		Darren Dahly	
5					4			red			
6											
7											
8											
9											
10											
11											
12											
13											
14											
15											
16											
17											
18											
19											
20											
21											
22											
23											
24											
25											
26											
27											
28											
29											
30											

**Bottom Navigation:** data | dictionary | values | +

# Less stress, more success

# The movement towards FAIR data



# SCIENTIFIC DATA

110110  
0111101  
110111110  
011101101

Amended: Addendum

OPEN

SUBJECT CATEGORIES

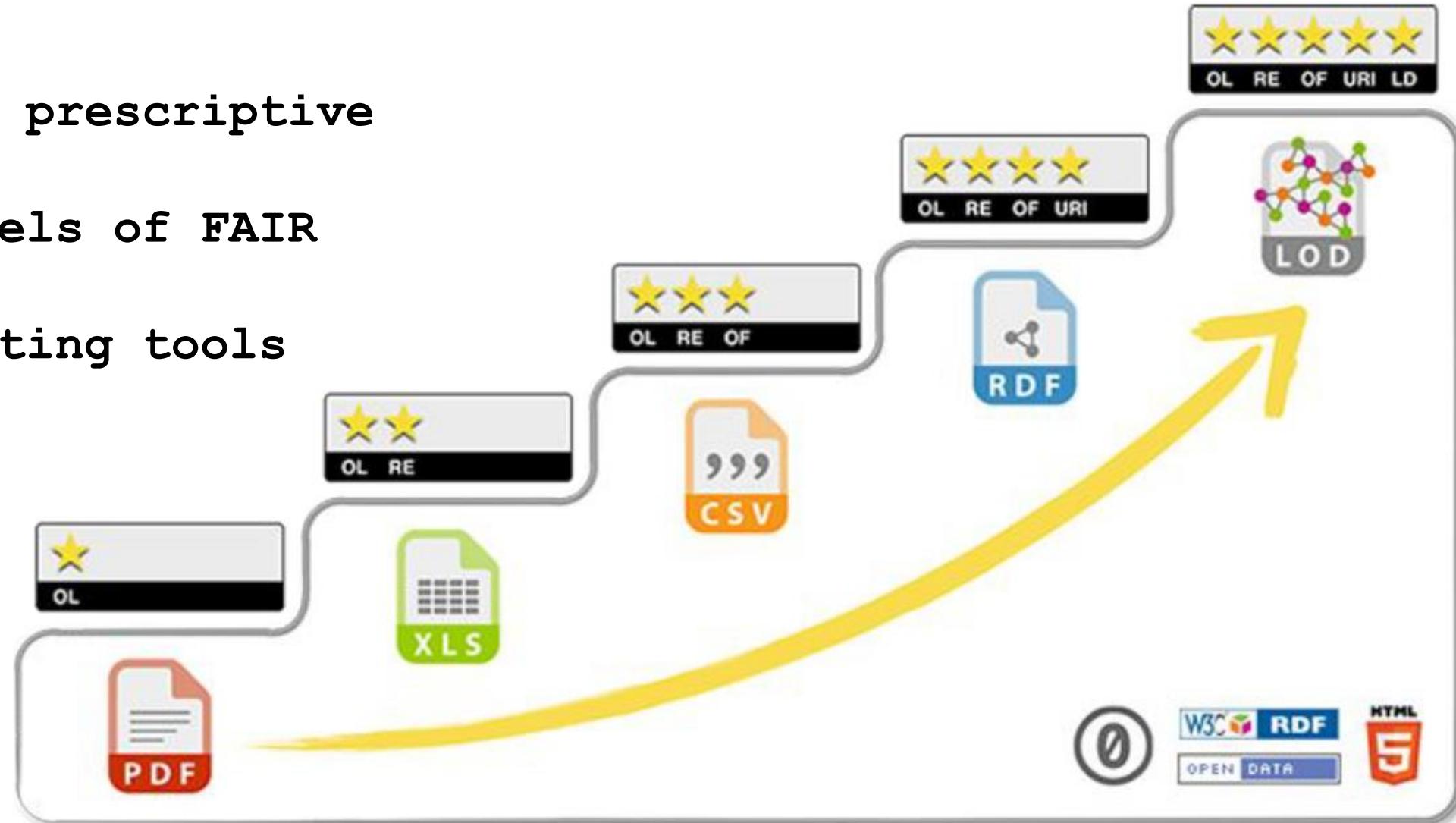
- » Research data
- » Publication characteristics

**Comment: The FAIR Guiding Principles for scientific data management and stewardship**

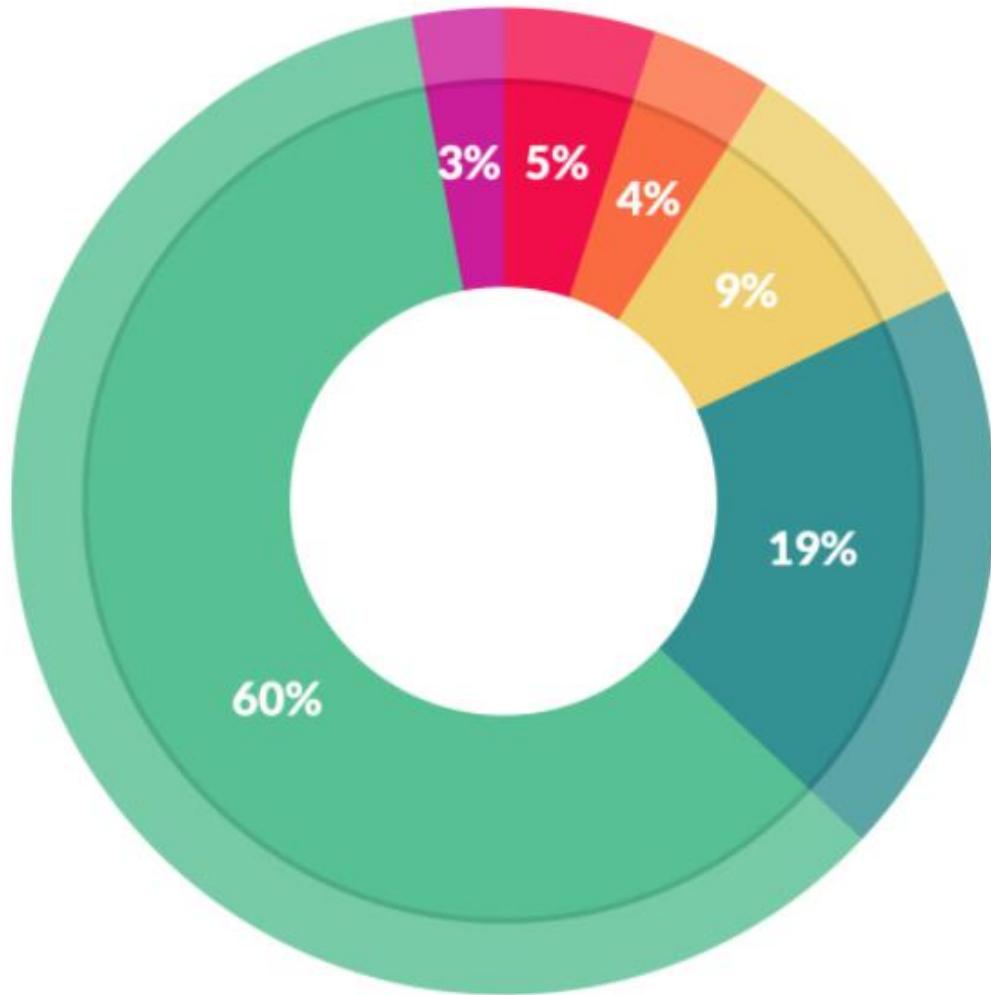
Mark D. Wilkinson *et al.* #

# Three main take home messages

- FAIR is not prescriptive
- Varying levels of FAIR
- Use of existing tools



# 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%*

# FAIR is a part of our life now!

## Data Guidelines

1. Background
  - 1.1 Open Data Policy
  - 1.2 Fair Data Principles

2. Share Your Data in 3 Steps
  - 2.1 Prepare Your Data for Sharing
  - 2.2 Select a Repository
  - 2.3 Add a Data Availability Statement to Your Manuscript
  - 2.4 Linking your datasets to your article

Some types of data benefit from visualization within the article. Wellcome Open Research welcomes the submission of manuscripts featuring [Plot.ly interactive figures](#) and [Code Ocean compute capsules](#). For further detail, please [contact us](#).



## Research Data Management

Good data governance and stewardship are key components of good research practice. In this regard, Science Foundation Ireland supports that research data should be Findable, Accessible, Interoperable and Reusable (FAIR)\*. Appropriate data management and data sharing are fundamental to all stages of the research process and support high quality, reproducible research. As such, access to research data arising in whole or in part from SFI funding should be as open as possible.



### FAIR Data Management

Describe the approach to data management that will be taken during and after the project, including who will be responsible for data management and data stewardship. The word limit is 500 words.



### Social Research Ethics Committee (SREC) ETHICS APPROVAL FORM

✉ [srec@ucc.ie](mailto:srec@ucc.ie)

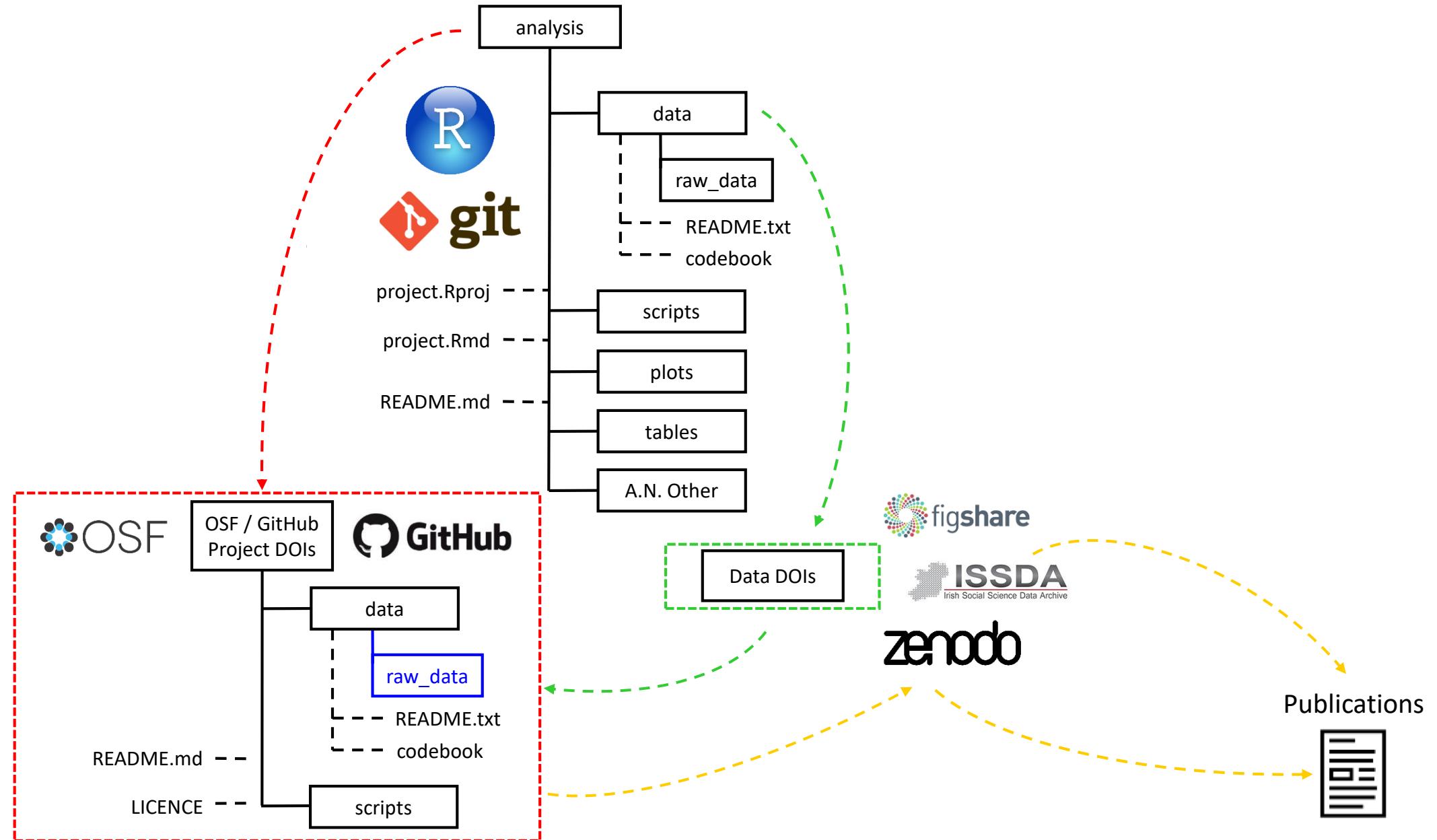
<https://www.ucc.ie/en/research/about/ethics/>

<sup>4</sup> Data management should follow the FAIR guiding principles (Findability, Accessibility, Interoperability & Reusability). See, for example, Wilkinson, M. D. et al. (2016) The FAIR Guiding Principles for Scientific Data Management and Stewardship. Full text: <http://www.nature.com/articles/sdata201618>. It is required that all staff and student researchers store those data which are required to replicate research findings, and the information required to enable re-use of data. Details of the UCC policy on research data storage can be found in section 8 of the Code of Research Conduct (2016): <https://www.ucc.ie/en/media/research/researchatucc/documents/UCCCodeofResearchConduct.pdf>. SREC advises against storing research data on non UCC approved cloud-based storage services. Physical data must be stored in a locked cabinet and you must specify who has permission to access this data.

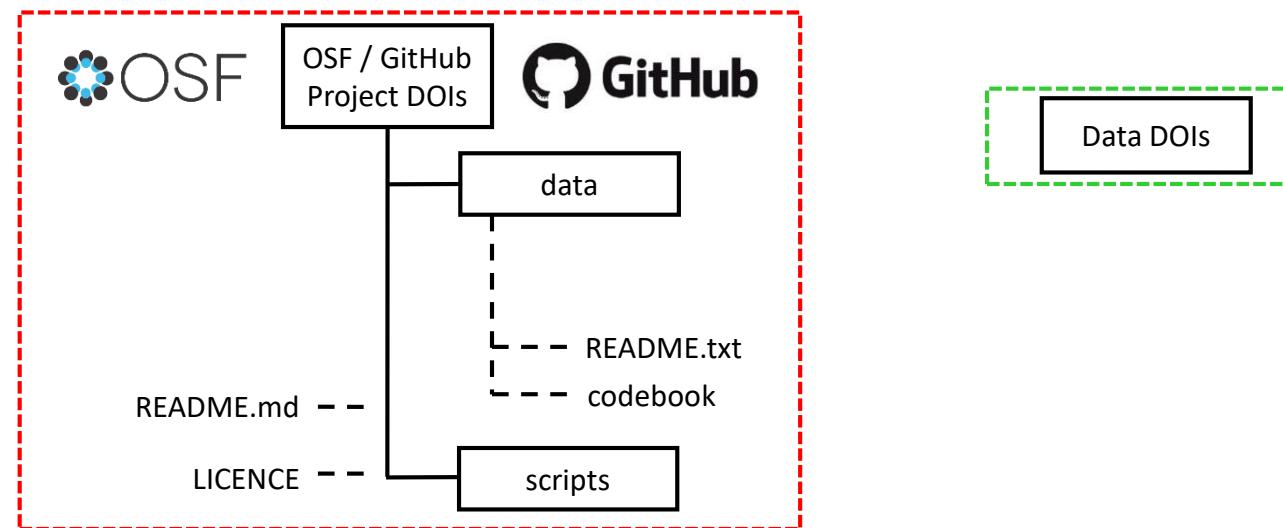


A set of Digital Object Compliance principles that describes the properties of digital objects that enables them to be findable, accessible, interoperable and reproducible (FAIR).

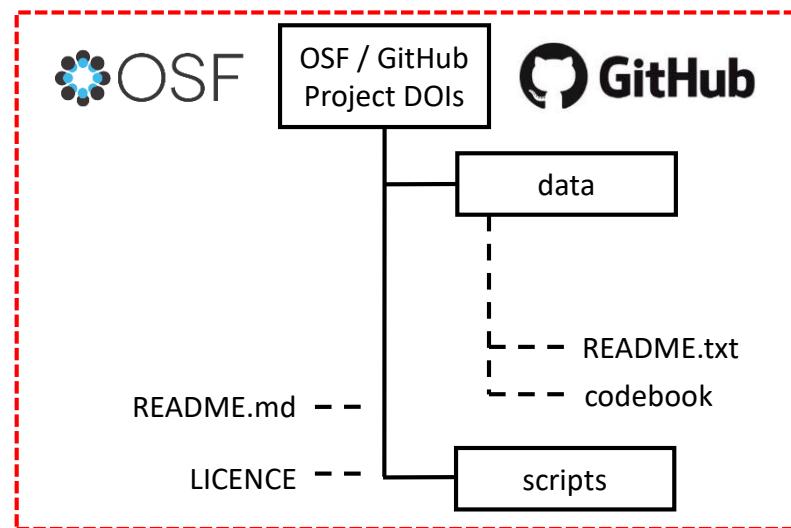
# What does this allow us to do?



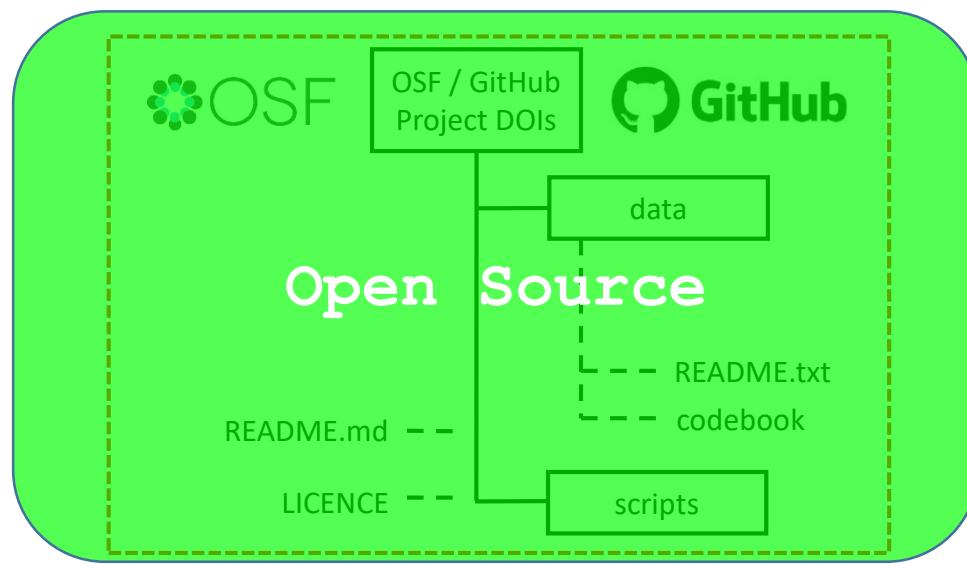
# What does this allow us to do?



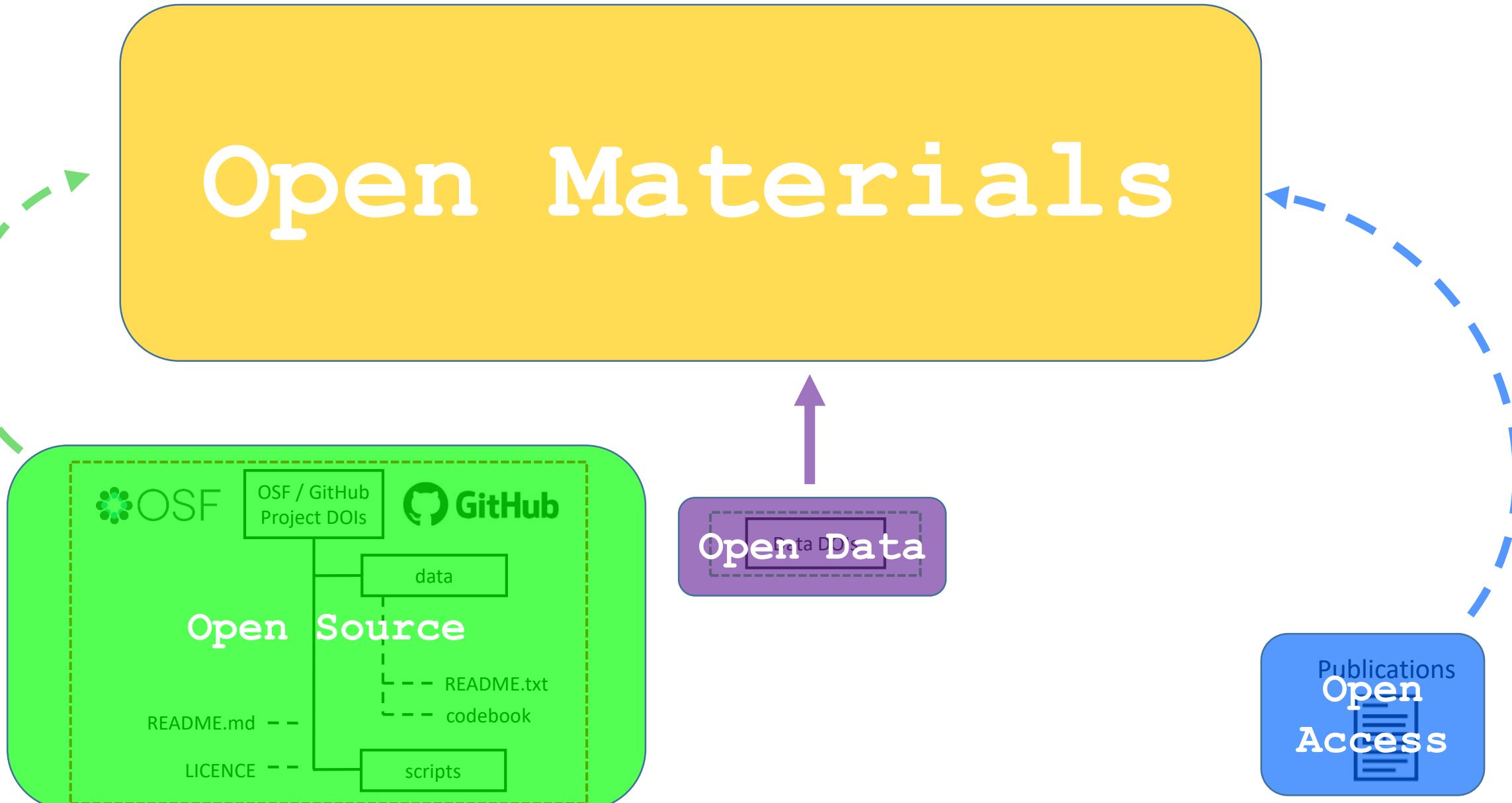
# What does this allow us to do?



# 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



**bioRxiv**  
THE PREPRINT SERVER FOR BIOLOGY

HOME | ABOUT | SUBMIT | ALERTS / RSS | CHANNELS

Search



Advanced Search

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

Download PDF

Email  
Share  
Citation Tools

Tweet Like 0

Subject Area

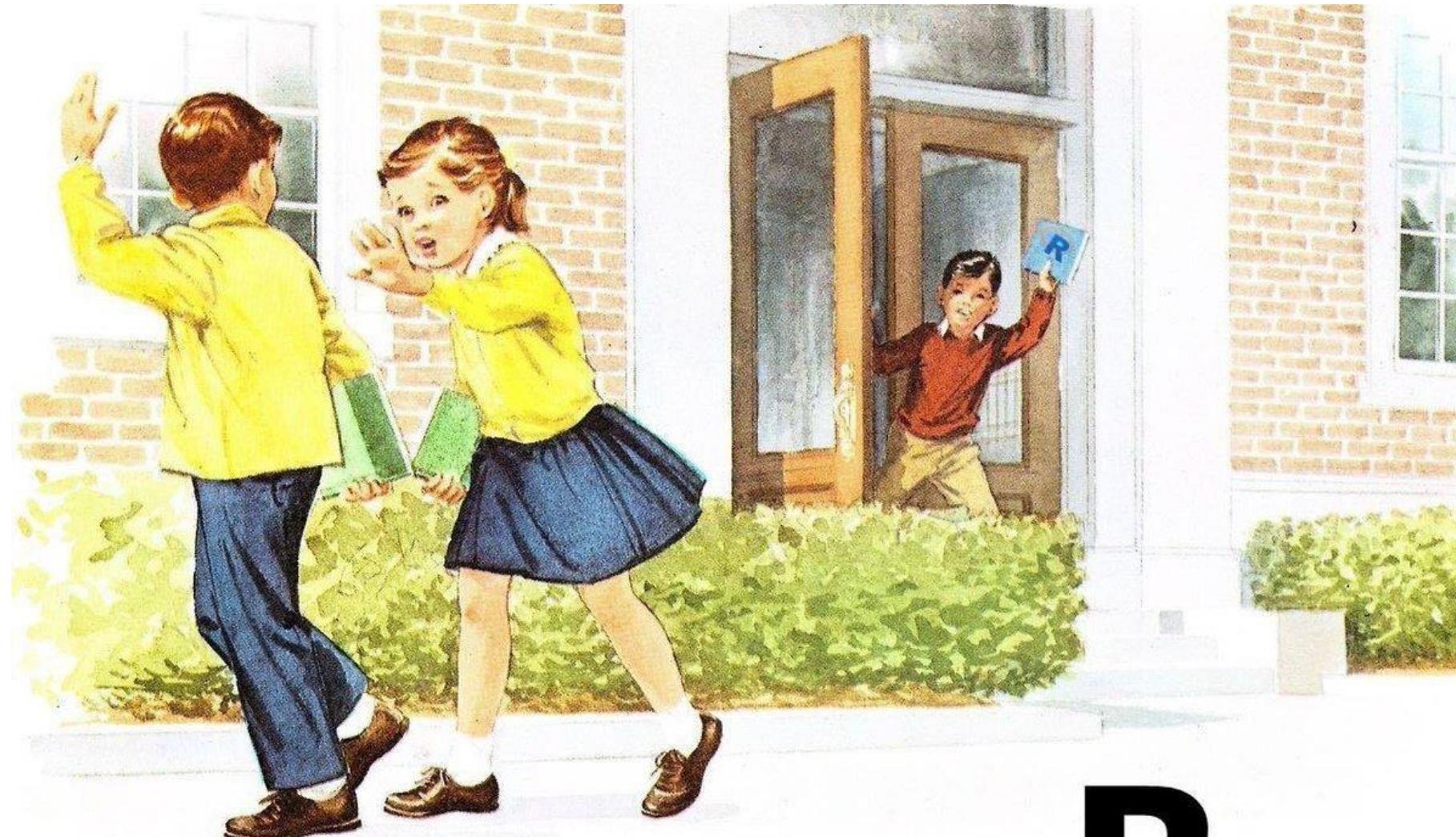
Genetics

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

# The key is they can be detected and corrected!



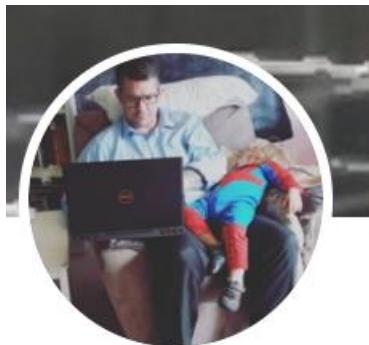
Run, or he's going to tell us about  
again!

R

# Acknowledgements



**Jenny Bryan**  
@JennyBryan



**Darren L Dahly**  
@statsepi



**Kara BOO!**   
@kara\_woo



**Karl Broman**  
@kwbroman



**Hadley Wickham**  
@hadleywickham



**UK Reproducibility  
Network**  
@ukrepro



**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.



**Dorothy Bishop**  
@deevybee



**Retraction Watch**  
@RetractionWatch