Computationally reproducible research

Leveraging reproducibility tools in laboratory-based research

Brooke Anderson, Colorado State University
Department of Environmental & Radiological Health Sciences

✉: brooke.anderson@colostate.edu
🐦: @gbwanderson
: github.com/geanders

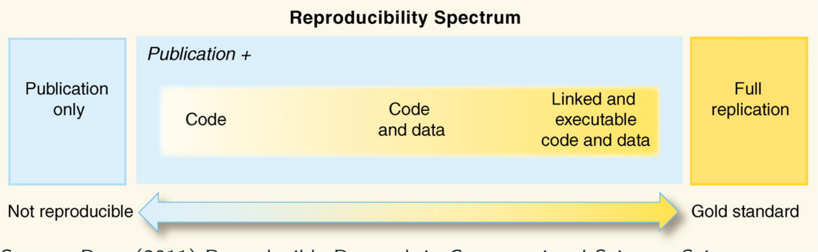
Objectives

The objectives for this talk are:

1. Clarify the principle and requirements for **reproducible research**, from a computational standpoint.
2. Outline some guidelines for **recording experimental data** in a way that facilitates computationally reproducible research, based on two recent papers:
   - Broman and Woo (2018) Data Organization in Spreadsheets, *The American Statistician*, 72:1, 2–10, DOI: 10.1080/00031305.2017.1375989
   - Ellis and Leek (2018) How to Share Data for Collaboration, *The American Statistician*, 72:1, 53–57, DOI: 10.1080/00031305.2017.1375987

Objective 1: Clarify the principle and requirements for reproducible research, from a computational standpoint.

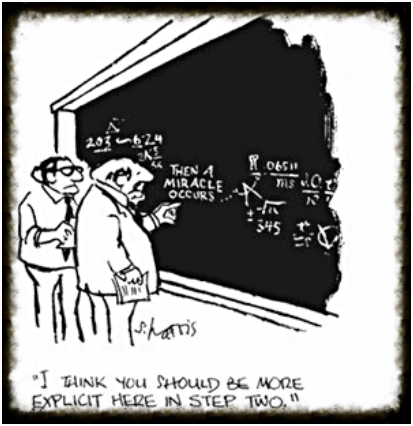
Reproducible research



Source: Peng (2011) Reproducible Research in Computational Science, *Science*, 334:6060, 1226–1227, DOI: 10.1126/science.1213847

Computationally **reproducible research** is research for which another person could take the published materials and recreate the same results from the same raw data.

Reproducible research



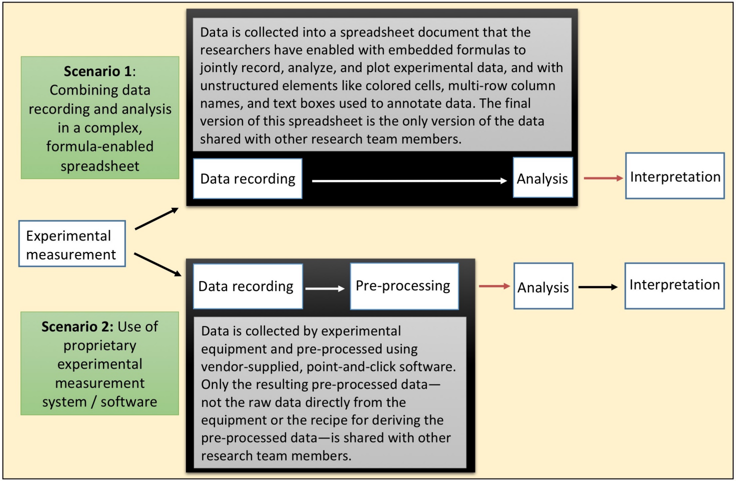
Source: Sidney Harris, The New Yorker

To make research computationally reproducible, full instructions should be available describing how you:

- Did any cleaning, pre-processing, or reformatting of the **raw data** (i.e., the data directly recorded for an experiment or output by laboratory equipment)
- Analyzed the **processed data** to generate figures, tables, and other research results

Code scripts are an excellent way to record this information.

Common “black boxes” in laboratory-based research



Scenario 1: Combining data recording and analysis in a complex, formula-enabled spreadsheet

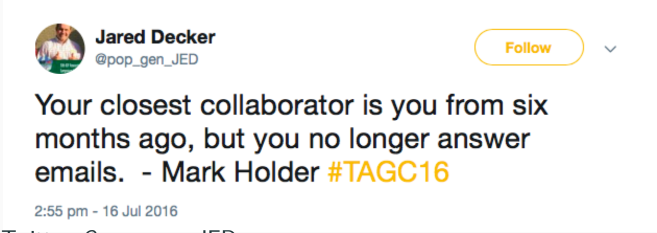
Data is collected into a spreadsheet document that the researchers have enabled with embedded formulas to jointly record, analyze, and plot experimental data, and with unstructured elements like colored cells, multi-row column names, and text boxes used to annotate data. The final version of this spreadsheet is the only version of the data shared with other research team members.

Scenario 2: Use of proprietary experimental measurement system / software

Data is collected by experimental equipment and pre-processed using vendor-supplied, point-and-click software. Only the resulting pre-processed data—not the raw data directly from the equipment or the recipe for deriving the pre-processed data—is shared with other research team members.

We identified two common **black boxes** in laboratory-based research, where the research steps are often neither **transparent** nor **reproducible**.

“Co-benefits” of reproducible research



Jared Decker
@pop_gen_JED

Your closest collaborator is you from six months ago, but you no longer answer emails. - Mark Holder #TAGC16

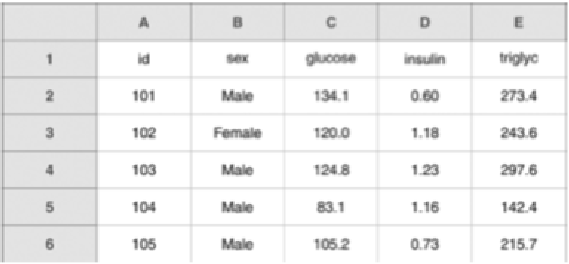
2:55 pm - 16 Jul 2016

Source: Twitter, @pop_gen_JED

Meeting the standards of reproducibility can have many co-benefits for a research lab, including **increasing efficiency** of research and **sharing data pre-processing and analysis techniques** across laboratory members.

Objective 2: Outline some guidelines for recording experimental data in a way that facilitates computationally reproducible research

Record data in “rectangular” formats



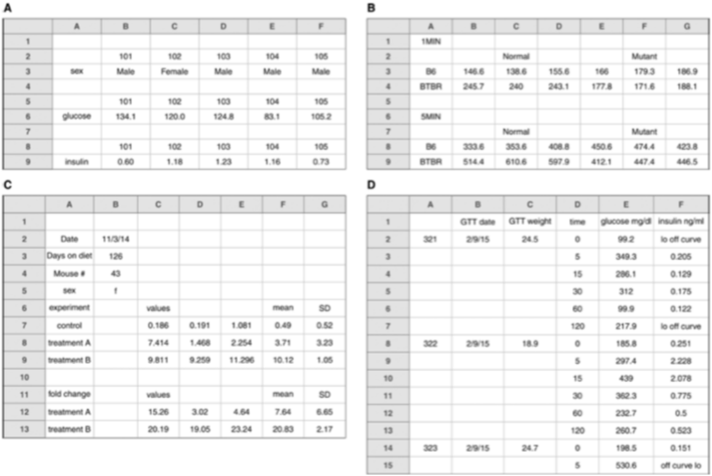
	A	B	C	D	E
1	id	sex	glucose	insulin	triglyc
2	101	Male	134.1	0.60	273.4
3	102	Female	120.0	1.18	243.6
4	103	Male	124.8	1.23	297.6
5	104	Male	83.1	1.16	142.4
6	105	Male	105.2	0.73	215.7

Figure 4. An example spreadsheet with a rectangular layout. This layout will aid future analyses.

Source: Broman and Woo, 2018

Rectangular format: One unit of observation per spreadsheet; one row for each study observation (e.g., study subject, time point); one column for each variable being measured; no empty boxes.

Non-“rectangular” formats



A

	A	B	C	D	E	F
1						
2		101	102	103	104	105
3	sex	Male	Female	Male	Male	Male
4						
5		101	102	103	104	105
6	glucose	134.1	120.0	124.8	83.1	105.2
7						
8		101	102	103	104	105
9	insulin	0.60	1.18	1.23	1.16	0.73

B

	A	B	C	D	E	F	G
1	1MIN						
2			Normal			Mutant	
3	B6	146.6	138.6	155.6	166	179.3	186.9
4	BTBR	245.7	240	243.1	177.8	171.6	188.1
5							
6	5MIN						
7			Normal			Mutant	
8	B6	333.6	353.6	408.8	450.6	474.4	423.8
9	BTBR	514.4	610.6	597.9	412.1	447.4	446.5

C

	A	B	C	D	E	F	G
1							
2	Date	11/3/14					
3	Days on diet	126					
4	Mouse #	43					
5	sex	f					
6	experiment		values			mean	SD
7	control		0.186	0.191	1.081	0.49	0.52
8	treatment A		7.414	1.468	2.254	3.71	3.23
9	treatment B		9.811	9.259	11.296	10.12	1.05
10							
11	fold change		values			mean	SD
12	treatment A		15.26	3.02	4.64	7.64	6.65
13	treatment B		20.19	19.05	23.24	20.83	2.17

D

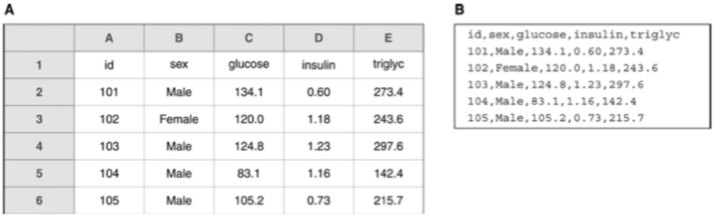
	A	B	C	D	E	F
1		GTT date	GTT weight	time	glucose mg/dl	insulin ng/ml
2	321	2/9/15	24.5	0	99.2	lo off curve
3				5	349.3	0.205
4				15	286.1	0.129
5				30	312	0.175
6				60	99.9	0.122
7				120	217.9	lo off curve
8	322	2/9/15	18.9	0	185.8	0.251
9				5	297.4	2.228
10				15	439	2.078
11				30	362.3	0.775
12				60	232.7	0.5
13				120	260.7	0.523
14	323	2/9/15	24.7	0	198.5	0.151
15				5	530.6	off curve lo

Figure 5. Examples of spreadsheets with nonrectangular layouts. These layouts are likely to cause problems in analysis.

Source: Broman and Woo, 2018

These may be **human-readable**, but are much less **computer-readable**.

Think in terms of “plain text” file formats



A

	A	B	C	D	E
1	id	sex	glucose	insulin	triglyc
2	101	Male	134.1	0.60	273.4
3	102	Female	120.0	1.18	243.6
4	103	Male	124.8	1.23	297.6
5	104	Male	83.1	1.16	142.4
6	105	Male	105.2	0.73	215.7

B

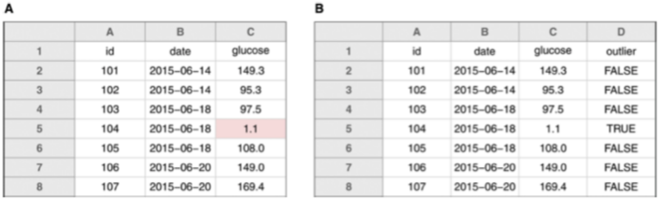
```
id,sex,glucose,insulin,triglyc
101,Male,134.1,0.60,273.4
102,Female,120.0,1.18,243.6
103,Male,124.8,1.23,297.6
104,Male,83.1,1.16,142.4
105,Male,105.2,0.73,215.7
```

Figure 11. (a) An example spreadsheet. (b) The same data as a plain text file in CSV format.

Source: Broman and Woo, 2018

Ideally, the data recording format should be something that could be set up as within a **plain text file format**, like a comma-separated values format (.csv).

Avoid cell formatting



A

	A	B	C
1	id	date	glucose
2	101	2015-06-14	149.3
3	102	2015-06-14	95.3
4	103	2015-06-18	97.5
5	104	2015-06-18	1.1
6	105	2015-06-18	108.0
7	106	2015-06-20	149.0
8	107	2015-06-20	169.4

B

	A	B	C	D
1	id	date	glucose	outlier
2	101	2015-06-14	149.3	FALSE
3	102	2015-06-14	95.3	FALSE
4	103	2015-06-18	97.5	FALSE
5	104	2015-06-18	1.1	TRUE
6	105	2015-06-18	108.0	FALSE
7	106	2015-06-20	149.0	FALSE
8	107	2015-06-20	169.4	FALSE

Figure 10. Highlighting in spreadsheets. (a) A potential outlier indicated by highlighting the cell. (b) The preferred method for indicating outliers, via an additional column.

Source: Broman and Woo, 2018

Any time you use **highlighting** or other forms of cell formatting in a spreadsheet, you will lose the information when you read the data into R or Python. Similarly, avoid adding **text boxes** or **embedded formulas** to spreadsheets used for data recording.

Be careful in naming columns

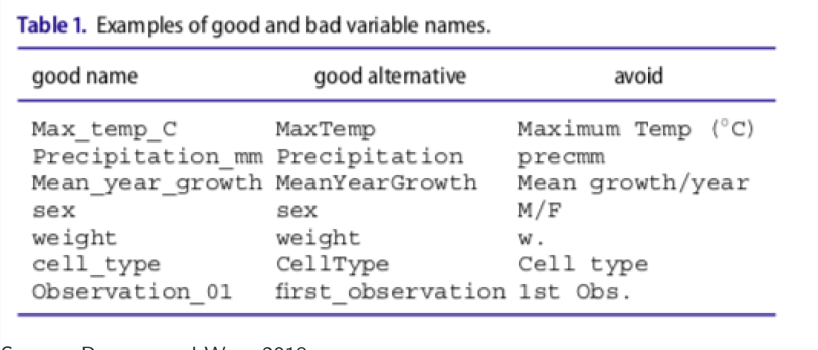


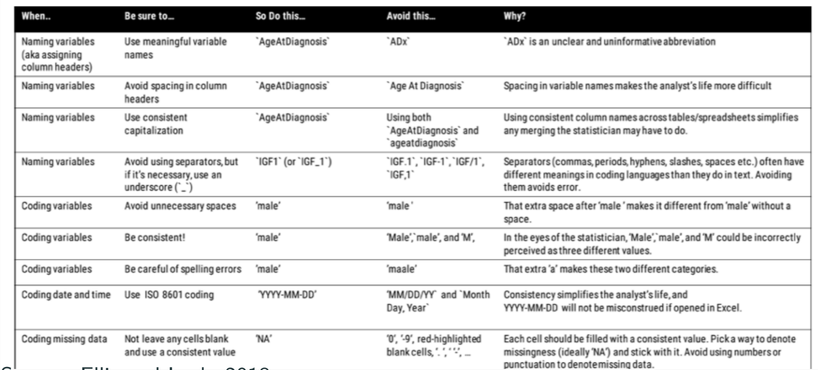
Table 1. Examples of good and bad variable names.

good name	good alternative	avoid
`Max_temp_C`	`MaxTemp`	`Maximum Temp (°C)`
`Precipitation_mm`	`Precipitation`	`precmm`
`Mean_year_growth`	`MeanYearGrowth`	`Mean growth/year`
`sex`	`sex`	`M/F`
`weight`	`weight`	`w.`
`cell_type`	`CellType`	`Cell type`
`Observation_01`	`first_observation`	`1st Obs.`

Source: Broman and Woo, 2018

Make sure that column names do not have **spaces**, **mathematical symbols**, or **other special characters**.

More guidelines



When..	Be sure to...	So Do this...	Avoid this...	Why?
Naming variables (aka assigning column headers)	Use meaningful variable names	`AgeAtDiagnosis`	`ADx`	`ADx` is an unclear and uninformative abbreviation
Naming variables	Avoid spacing in column headers	`AgeAtDiagnosis`	`Age At Diagnosis`	Spacing in variable names makes the analyst's life more difficult
Naming variables	Use consistent capitalization	`AgeAtDiagnosis`	Using both `AgeAtDiagnosis` and `ageatdiagnosis`	Using consistent column names across tables/spreadsheets simplifies any merging the statistician may have to do.
Naming variables	Avoid using separators, but if it's necessary, use an underscore (`_`)	`IGF1` (or `IGF_1`)	`IGF.1`, `IGF-1`, `IGF/1`, `IGF,1`	Separators (commas, periods, hyphens, slashes, spaces etc.) often have different meanings in coding languages than they do in text. Avoiding them avoids error.
Coding variables	Avoid unnecessary spaces	'male'	'male '	That extra space after 'male ' makes it different from 'male' without a space.
Coding variables	Be consistent!	'male'	'Male','male', and 'M',	In the eyes of the statistician, 'Male','male', and 'M' could be incorrectly perceived as three different values.
Coding variables	Be careful of spelling errors	'male'	'maale'	That extra 'a' makes these two different categories.
Coding date and time	Use ISO 8601 coding	'YYYY-MM-DD'	'MM/DD/YY' and `Month Day, Year`	Consistency simplifies the analyst's life, and YYYY-MM-DD will not be misconstrued if opened in Excel.
Coding missing data	Not leave any cells blank and use a consistent value	'NA'	'0', '-9', red-highlighted blank cells, '. ', ' '-', ...	Each cell should be filled with a consistent value. Pick a way to denote missingness (ideally 'NA') and stick with it. Avoid using numbers or punctuation to denote missing data.

Source: Ellis and Leek, 2018

Similar and additional guidelines are outlined in Ellis and Leek, 2018.