

# Data management

Fall 2017

Dr. Philipp Boersch-Supan

[pboesu@gmail.com](mailto:pboesu@gmail.com)

Slides at [pboesu.github.io/teaching](https://pboesu.github.io/teaching)

Why do we care about good  
**data management**  
and  
**reproducibility?**

**Funder requirement?**

# NSF, NIH, and many other funders require data management plans

Data Management Planning Tool: <https://dmptool.org/>

**Journal requirement?**

An increasing number of journals require deposition of data in public repositories as a condition of publishing

E.g. all ESA and BES journals, *Biology Letters*, *Proceedings B*, *Ecology and Evolution*...

**Noble cause?**

Opening up data sources and  
analysis code should make  
science more transparent and  
reproducible



**Egotism!**

**You will have to redo most analyses**

- because you have new data
- because you made a mistake
- because your advisor wants to change something
- because your committee wants a change
- because reviewer #2 wants a change
- because reviewer #2 didn't want the change anyways
- because every journal has different formatting requirements
- etc
- etc

# Computational effort

1. design study easy
2. collect data easy
3. clean & format data hard
4. descriptive statistics easy
5. inferential statistics moderate
6. reporting moderate - hard

**Start good data and analysis  
habits now**

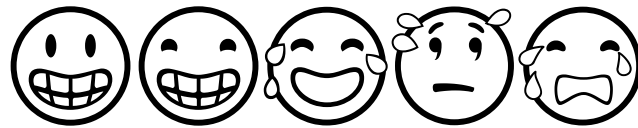
# Organisation starts on your computer

```
-- project_folder
|-- CITATION
|-- README
|-- LICENSE
|-- requirements.txt
|-- data
|   -- birds_count_table.csv
|-- doc
|   -- notebook.md
|   -- manuscript.md
|   -- changelog.txt
|-- results
|   -- summarized_results.csv
|-- src
|   -- sightings_analysis.py
|   -- runall.py
```

# Resources:

- A Quick Guide to Organizing Computational Biology Projects
  - <https://doi.org/10.1371/journal.pcbi.1000424>
- British Ecological Society Data Management Guide
  - <http://www.britishecologicalsociety.org/publications/guides-to/>
- Ten Simple Rules for Creating a Good Data Management Plan
  - <https://doi.org/10.1371/journal.pcbi.1004525>

# Spreadsheets





# Spreadsheets are not going away

- ~1 billion use Microsoft Office
- ~650 million use spreadsheets
- >50% use formulas
- 250K - 1 million people use R

Spreadsheets make it easy to  
enter data

# Spreadsheets combine

- data
- figures
- formatting
- interactive calculations
- reactive properties (e.g. “smart” formatting)

often in proprietary data formats

Spreadsheets make it easy to  
make a mess of data

Use spreadsheets to prepare  
*machine readable data.*

Don't use spreadsheets to do  
your analysis.

# Excel analysis vs R analysis

- easy to get started
  - easy to be inconsistent
  - no logical sequence to your actions
  - hard to document your actions
  - tweak a detail: usually lots of manual work
  - reproducibility limited to impossible
- hard to start
  - inconsistencies can break analysis
  - scripts have an order → sequence of analysis steps
  - easy to document your actions with comments
  - tweak a detail: usually small change in code
  - great for reproducibility

# Nope

	C	D
1		
2	Comment	
3		
4	In ethonal	
5	In ethonal	
6	In ethanol	
7	In plastic bag. Frozen	
8	IN plastic bag	
9		



# Be consistent

	C	D
1		
2	Comment	
3		
4	In ethonal	
5	In ethonal	
6	In ethanol	
7	In plastic bag. Frozen	
8	IN plastic bag	
9		

# Nope

	A	B	C	D	E
32	59	22	1	grass	0
33	40	75	1	M	1
34	64	53	1	Bare/forest	0
35	34	55	1	Bare/forest	0
36	35	72	1	M	0
37	36	38	1	M	0
38	45	47	1	M/grass	0
39	41	44	1	Bushes	0
40	42	30	1	M	0
41	44	25	1	M	0
42	60	50	1	M	0
43	65	25	1	Grass	0
44					<b>0.2195122</b>
45	Average	58.658537			
46	s.e.	22.605876			

# Put just one thing in a cell

	A	B	C	D	E
32	59	22	1	grass	0
33	40	75	1	M	1
34	64	53	1	Bare/forest	0
35	34	55	1	Bare/forest	0
36	35	72	1	M	0
37	36	38	1	M	0
38	45	47	1	M/grass	0
39	41	44	1	Bushes	0
40	42	30	1	M	0
41	44	25	1	M	0
42	60	50	1	M	0
43	65	25	1	Grass	0
44					<b>0.2195122</b>
45	Average	58.658537			
46	s.e.	22.605876			

# Nope

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
1		7								9						
2	Date	Date	Sticks	Weight (g)	Bill (mm)	Tarsus (mm)	Wing (mm)		Date	Sticks	Weight (g)	Bill (mm)	Tarsus (mm)	Wing (mm)		
3	22-Oct	22-Oct							22-Oct							
4	23-Oct	23-Oct							23-Oct							
5	24-Oct	24-Oct							24-Oct							
6	25-Oct	25-Oct							25-Oct	Down	40	12.8	18.7	23		
7	26-Oct	26-Oct							26-Oct	Down	55					
8	27-Oct	27-Oct							27-Oct	Down	66					
9	28-Oct	28-Oct							28-Oct	Down	71					
10	29-Oct	29-Oct							29-Oct	Down	75					
11	30-Oct	30-Oct							30-Oct	Down	70					
12	31-Oct	31-Oct							31-Oct	Down	88	14.7	22.9	27		
13	1-Nov	1-Nov							1-Nov	Down	110					
14	2-Nov	2-Nov	Down	43	22.1	18.0	22.1		2-Nov	Down	130					
15	3-Nov	3-Nov	Down	58					3-Nov	Down	123					
16	4-Nov	4-Nov	Down	55					4-Nov	Down	115					
17	5-Nov	5-Nov	Down	63					5-Nov	Down	148					
18	6-Nov	6-Nov	Down	62					6-Nov	Down	157	17.9	28.7	37		
19	7-Nov	7-Nov	Down	113	15.8	22.4	27		7-Nov	Up	163					
20	8-Nov	8-Nov	Down	90					8-Nov	Down	160					
21	9-Nov	9-Nov	Down	120					9-Nov	Down	180					
22	10-Nov	10-Nov	Down	133					10-Nov	Down	188					
23	11-Nov	11-Nov	Down	135					11-Nov	Down	180					
24	12-Nov	12-Nov	Down	145					12-Nov	Down	160	21.3	34	43		
25	13-Nov	13-Nov							13-Nov							
26																
27																
28	11-Dec	11-Dec	Down	338	29.5	45.9	114		11-Dec	Down	358	28.9	41.5	131		
29	12-Dec	12-Dec	Down	320					12-Dec	Down	388					
30	13-Dec	13-Dec	Down	340					13-Dec	Down	388					
31	14-Dec	14-Dec	Down	356					14-Dec	Down	408					
32	15-Dec	15-Dec	Down	328					15-Dec	Down	378					
33	16-Dec	16-Dec	Down	328					16-Dec	Down	378					

# Make it a rectangle

# No empty cells

# No special characters in variable names

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O		
1		7								9							
2	Date	Date	Sticks	Weight (g)	Bill (mm)	Tarsus (mm)	Wing (mm)		Date	Sticks	Weight (g)	Bill (mm)	Tarsus (mm)	Wing (mm)			
3	22-Oct	22-Oct							22-Oct								
4	23-Oct	23-Oct							23-Oct								
5	24-Oct	24-Oct							24-Oct								
6	25-Oct	25-Oct							25-Oct	Down	40	12.8	18.7	23			
7	26-Oct	26-Oct							26-Oct	Down	55						
8	27-Oct	27-Oct							27-Oct	Down	66						
9	28-Oct	28-Oct							28-Oct	Down	71						
10	29-Oct	29-Oct							29-Oct	Down	75						
11	30-Oct	30-Oct							30-Oct	Down	70						
12	31-Oct	31-Oct							31-Oct	Down	88	14.7	22.9	27			
13	1-Nov	1-Nov							1-Nov	Down	110						
14	2-Nov	2-Nov	Down	43	22.1	18.0	22.1		2-Nov	Down	130						
15	3-Nov	3-Nov	Down	58					3-Nov	Down	123						
16	4-Nov	4-Nov	Down	55					4-Nov	Down	115						
17	5-Nov	5-Nov	Down	63					5-Nov	Down	148						
18	6-Nov	6-Nov	Down	62					6-Nov	Down	157	17.9	28.7	37			
19	7-Nov	7-Nov	Down	113	15.8	22.4	27		7-Nov	Up	163						
20	8-Nov	8-Nov	Down	90					8-Nov	Down	160						
21	9-Nov	9-Nov	Down	120					9-Nov	Down	180						
22	10-Nov	10-Nov	Down	133					10-Nov	Down	188						
23	11-Nov	11-Nov	Down	135					11-Nov	Down	180						
24	12-Nov	12-Nov	Down	145					12-Nov	Down	160	21.3	34	43			
25	13-Nov	13-Nov							13-Nov								
26																	
27																	
28	11-Dec	11-Dec	Down	338	29.5	45.9	114		11-Dec	Down	358	28.9	41.5	131			
29	12-Dec	12-Dec	Down	320					12-Dec	Down	388						
30	13-Dec	13-Dec	Down	340					13-Dec	Down	388						
31	14-Dec	14-Dec	Down	356					14-Dec	Down	408						
32	15-Dec	15-Dec	Down	328					15-Dec	Down	378						
33	16-Dec	16-Dec	Down	328					16-Dec	Down	378						





# No, no, no, no, no!

	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE
1	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15	D16	D17	D18	D19	D20
2		*				*	*	*												
3		*				*	*	*												
4		*				*	*	*		*	*									
5				*	*	*														
6		*			*	*	*													
7		*			*	*	*													
8				*	*	*														
9		*			*	*	*													
10		*			*	*	*													
11				*	*	*														
12		*		*	*	*														
13		*		*	*	*														
14				*	*	*														
15				*	*	*														
16				*	*	*														
17			*	*	*	*														
18			*	*	*	*														
19			*	*	*	*														
20			*	*	*	*														
21			*	*	*	*														
22			*	*	*	*														
23						*	*	*												
24																				
25						*	*	*												
26					*	*	*													
27																				
28					*	*	*													
29																				
30																				
31																				
32																				
33																				
34																				
35																				
36																				
37																				
38	*																			
39																				
40																				

Start  
 Alive  
 Dead  
 Alive  
 \* destructive sampling event



# Do not use font or cell color as data

	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE
1	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15	D16	D17	D18	D19	D20
2	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
3	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
4	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
5	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
6	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
7	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
8	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
9	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
10	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
11	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
12	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
13	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
14	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
15	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
16	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
17	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
18	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
19	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
20	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
21	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
22	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
23	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
24	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
25	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
26	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
27	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
28	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
29	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
30	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
31	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
32	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
33	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
34	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
35	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
36	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
37	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
38	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
39	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
40	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*

Legend:

- Start
- Alive
- Dead
- Alive
- \* destructive sampling event

# three principles for names

- machine readable
- human readable
- plays well with default ordering

# File names

- **BAD**

- myabstract.docx
- Joe's Filenames Use Spaces and Punctuation.xlsx
- figure 1.png
- fig 2.png
- JW7d^(2sl@deletethisandyourcareerisoverWx2\*.txt

- **GOOD**

- 2014-06-08\_abstract-for-sla.docx
- joes-filenames-are-getting-better.xlsx
- fig01\_scatterplot-talk-length-vs-interest.png
- fig02\_histogram-talk-attendance.png
- 1986-01-28\_raw-data-from-challenger-o-rings.txt

# 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.

---

# “machine readable”

- Search friendly
- Avoid
  - Spaces
  - Punctuation
  - Symbols, accented characters
  - case sensitivity
- easy to compute on
  - deliberate use of delimiters like “-” and “\_”

# “human readable”

- name contains info on ***content***

# “human readable”

- 01\_marshal-data.md
- 01\_marshal-data.r
- 02\_pre-dea-filtering.md
- 02\_pre-dea-filtering.r
- 03\_dea-with-limma-voom.md
- 03\_dea-with-limma-voom.r
- 04\_explore-dea-results.md
- 04\_explore-dea-results.r
- 90\_limma-model-term-name-fiasco.md
- 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
- 01.md
- 01.r
- 02.md
- 02.r
- 03.md
- 03.r
- 04.md
- 04.r
- 90.md
- 90.r
- helper01.r
- helper02.r
- helper03.r
- helper04.r

Which set of file(name)s do you want at 3a.m. before a deadline?

# “plays well with default ordering”

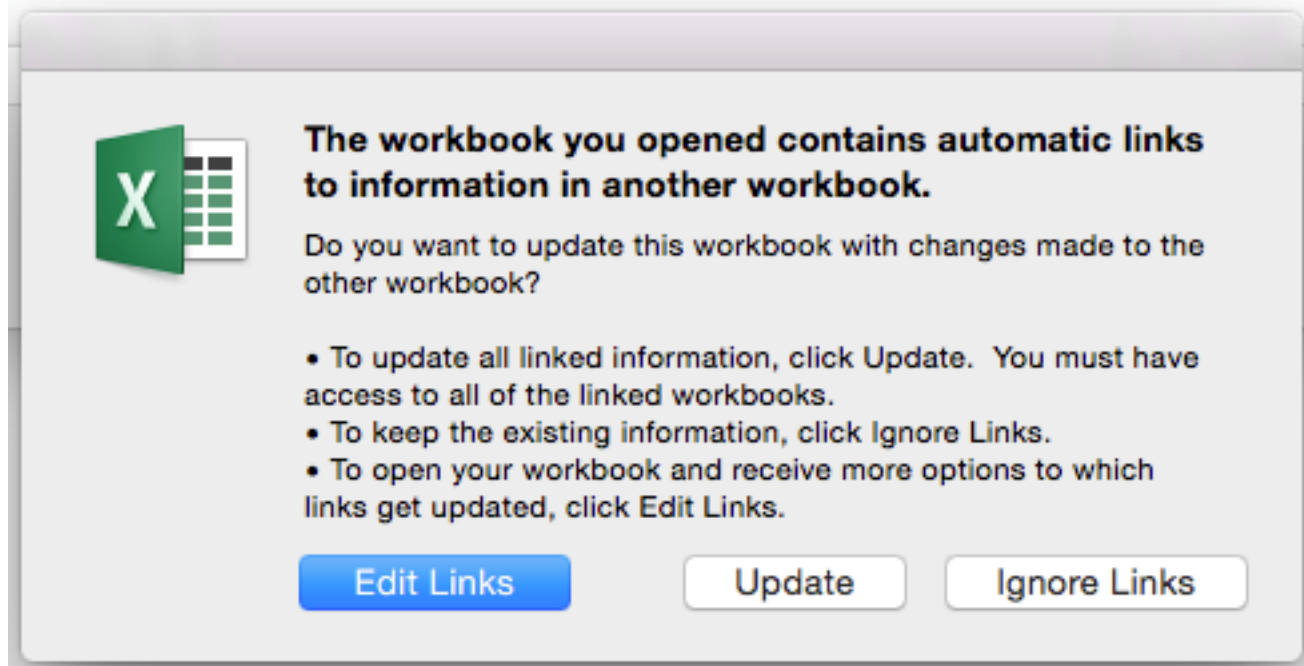
- put something numeric first
  - Sequence number for **logical ordering**
  - Date/timestamp for **chronological ordering**
- use the ISO 8601 standard for dates
  - YYYY-MM-DD
- left pad other numbers with zeros

10\_final-figs-for-publication.R  
1\_data-cleaning.R  
2\_fit-model.R

01\_data-cleaning.R  
02\_fit-model.R  
10\_final-figs-for-publication.R

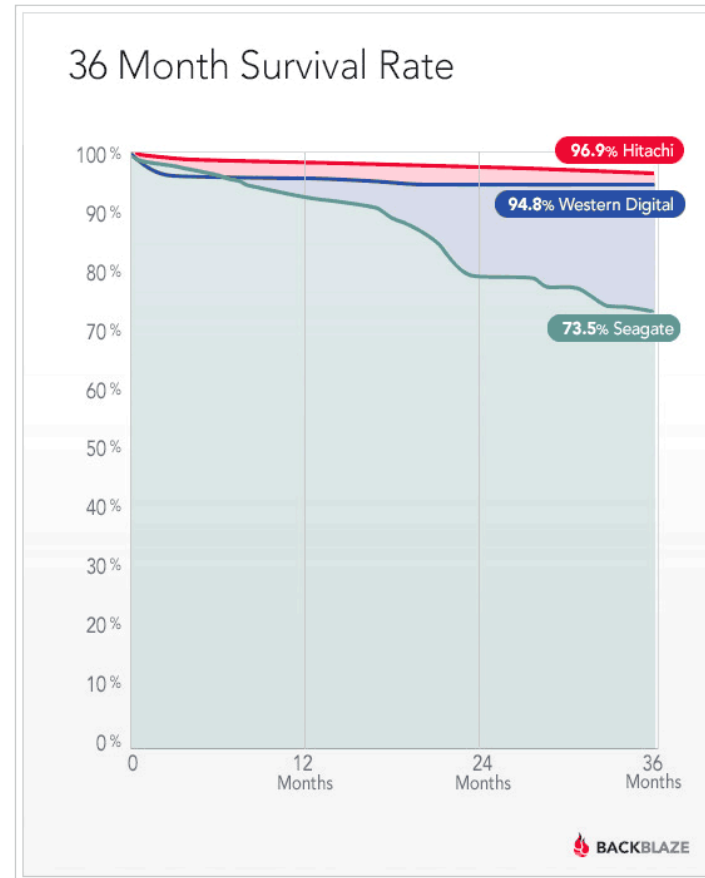
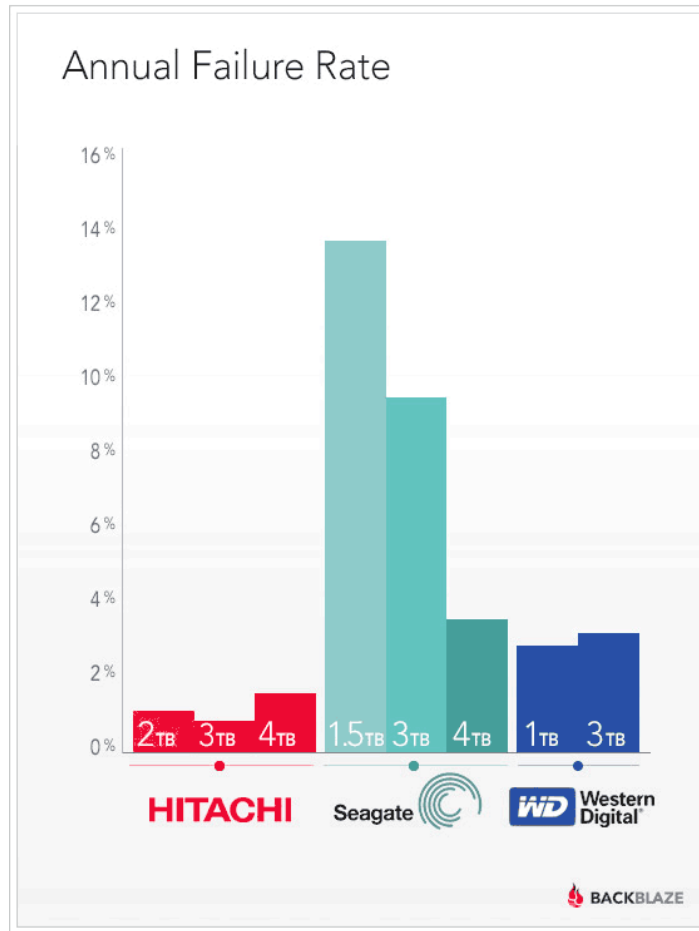


# Also this 🔥 🔥 🔥



- Excel offers many features that can trip you up later.
- Use individual sheets for individual datasets.
- Do the crosslinking in your analysis scripts.

# Make backups



- Most likely duration of a STEM Ph.D. in the US: 6 years

# Recap: spreadsheets

- Be consistent.
- Write dates as YYYY-MM-DD.
- Fill in all of the cells.
- Put just one thing in a cell.
- Make it a rectangle.
- Keep metadata and/or a data dictionary.
- No calculations in the raw data files.
- Don't use font color or highlighting as data.
- Choose good names for things.
- Use data validation to avoid data entry mistakes.
- Save the data in plain text files.
- Make backups.

# Resources:

- <http://kbroman.org/dataorg/>
- Data organization in spreadsheets
  - <https://doi.org/10.7287/peerj.preprints.3183v1>

# Where to get more training?

- **The Carpentries** teach researchers computing skills they need to get more done in less time & with less pain.
- Data wrangling skills
  - Spreadsheets, OpenRefine, SQL, R, Python
  - <http://www.datacarpentry.org/workshops-upcoming/>
- Software/programming skills
  - R, SQL, Python, Unix shell, version control
  - <https://software-carpentry.org/workshops/>
- Usually several workshops per year at UF in Gainesville
- Your lab/BGSO/etc. can organize a workshop locally
  - <https://software-carpentry.org/workshops/request/>