Data Organization in spreadsheets

Tobin Magle, PhD

2022-07-12

10:00 am Central

Workshop Materials

https://github.com/nuitrcs/data-org-spreadsheets

inspired by the Data Carpentry Ecology lesson
https://datacarpentry.org/spreadsheet-ecology-lesson/

Main questions

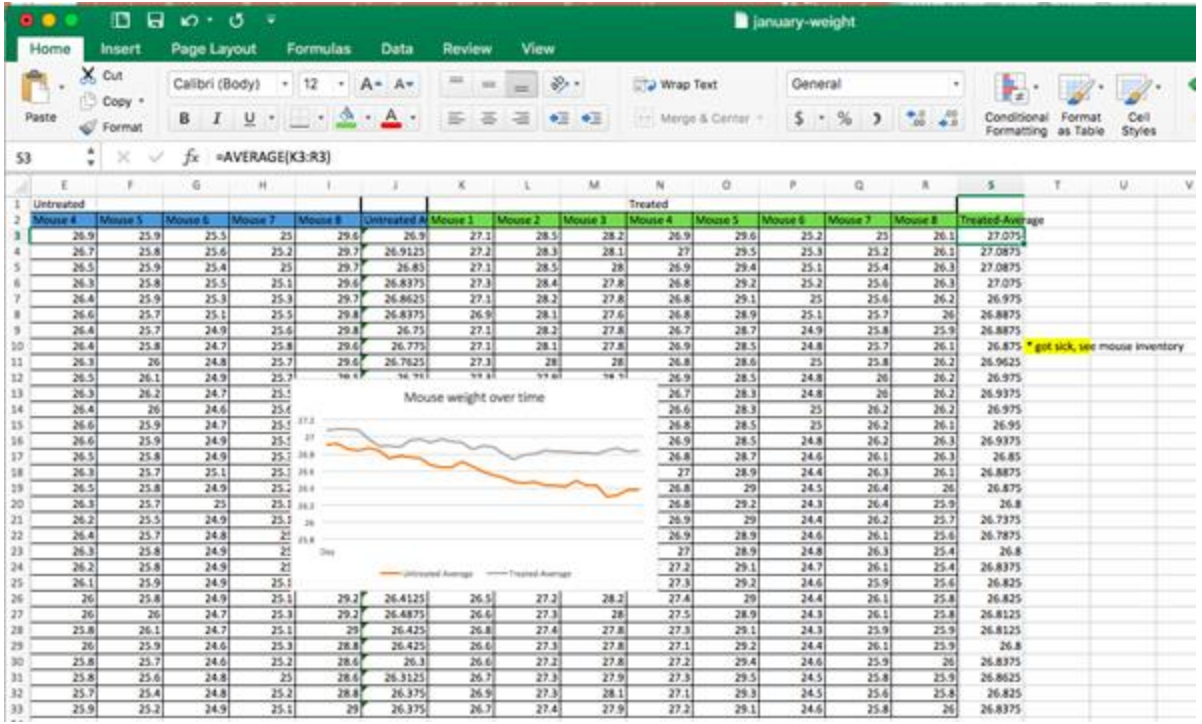
- How do you currently use spreadsheets?
  - Use a sample dataset with common scenarios
- How do computers read spreadsheets?
  - Necessary for automation
- How can you use spreadsheets to make them both machine and human readable?

Poll: how do you use spreadsheets?

- **Q1**: Do you use spreadsheets in your research?
  - In Chat: What do you use spreadsheets for?
- **Q2**: Have you have ever done something to your spreadsheet data that has made you frustrated or sad?
  - In Chat: What was it?

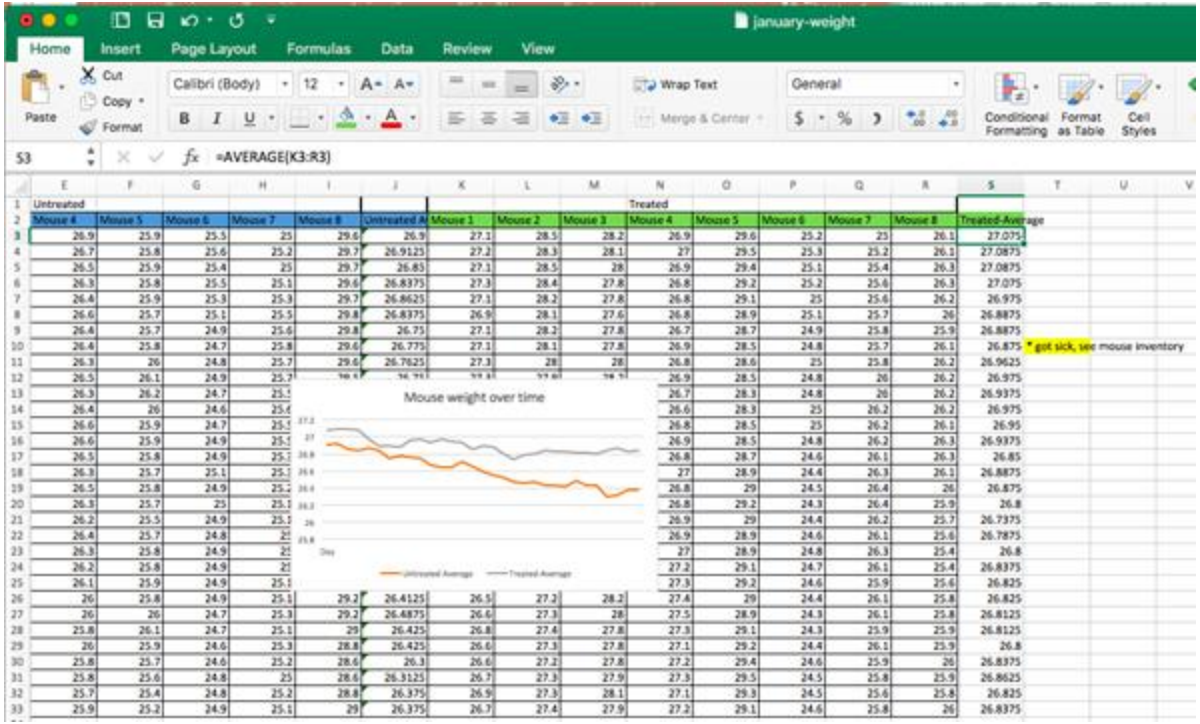
Common spreadsheet features

- Data
- Notes
- Formatting
- Color coding
- Many tabs
- Calculations
- Graphs



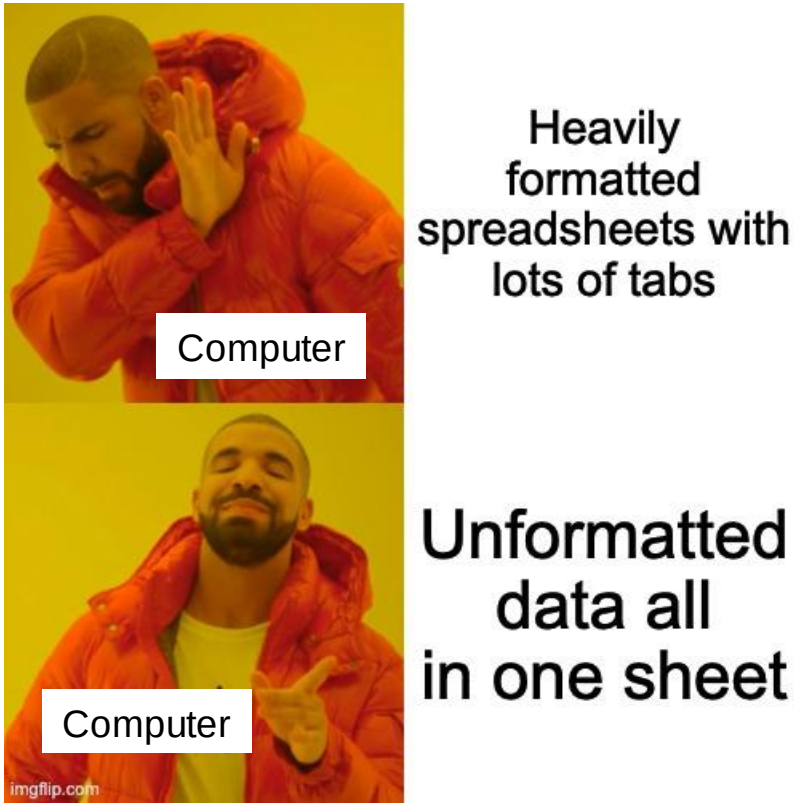
Pros: Human readable, all in one place

- Data
- Notes
- Formatting
- Color coding
- Many tabs
- Calculations
- Graphs



Con: Not machine-readable

- Data
- ~~Notes~~
- ~~Formatting~~
- ~~Color coding~~
- ~~Many tabs~~
- ~~Calculations~~
- ~~Graphs~~

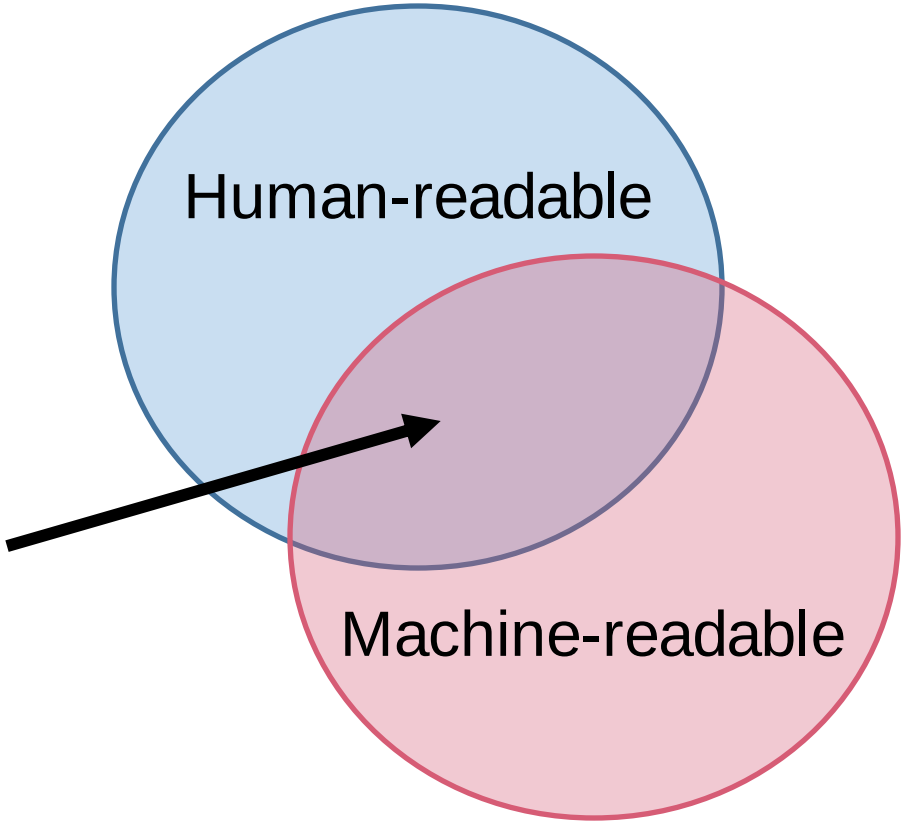


Goal

Take a **human readable** spreadsheet

to

A **machine readable** table that can be used in R/Python



Sample data: a survey of small mammals

- **Rows**: observations of individual animals
- **Columns**: Variables that describe the animals
  - Species, sex, date, location, etc
- Inconsistencies in data collection

Download the sample data

https://github.com/nuitrcs/data-org-spreadsheets

- Open the file in your spreadsheet program

How does a computer read...?

- **Multiple tabs**
- **Formatting**
- **Multiple tables**
- **Columns**
- **Missing data**

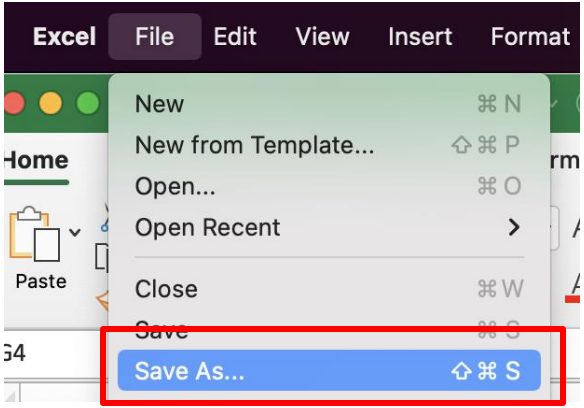
How do I know what a computer can read?

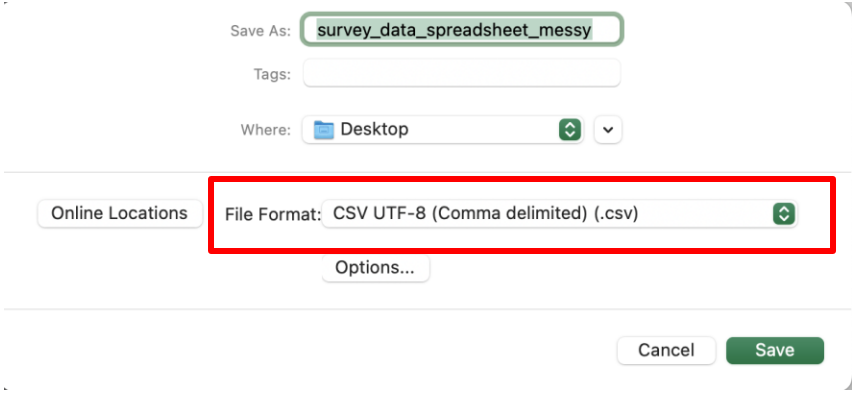
Save data as .csv

- R/Python read text-based file formats best
- To get an idea of what R and python will see, save as .csv

Demo: Save to .csv

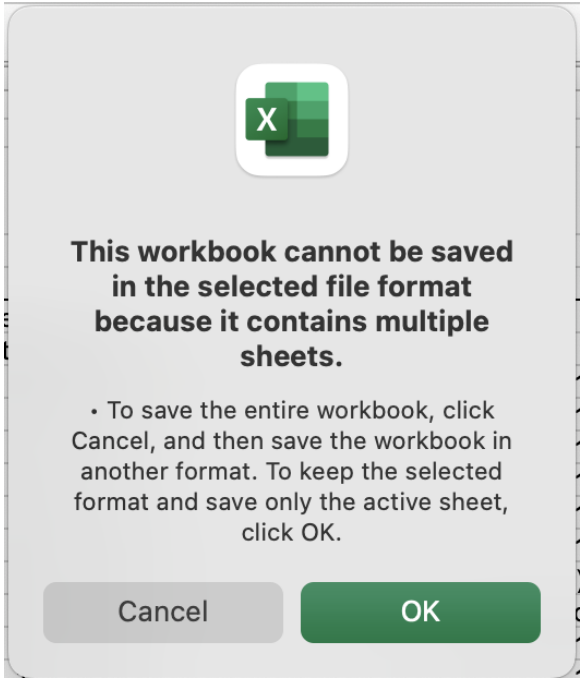
- Open the 2014 tab
- File > Save as
- File Format: csv





Demo: Save to .csv

- Open the 2014 tab
- File > Save as
- File Format: csv

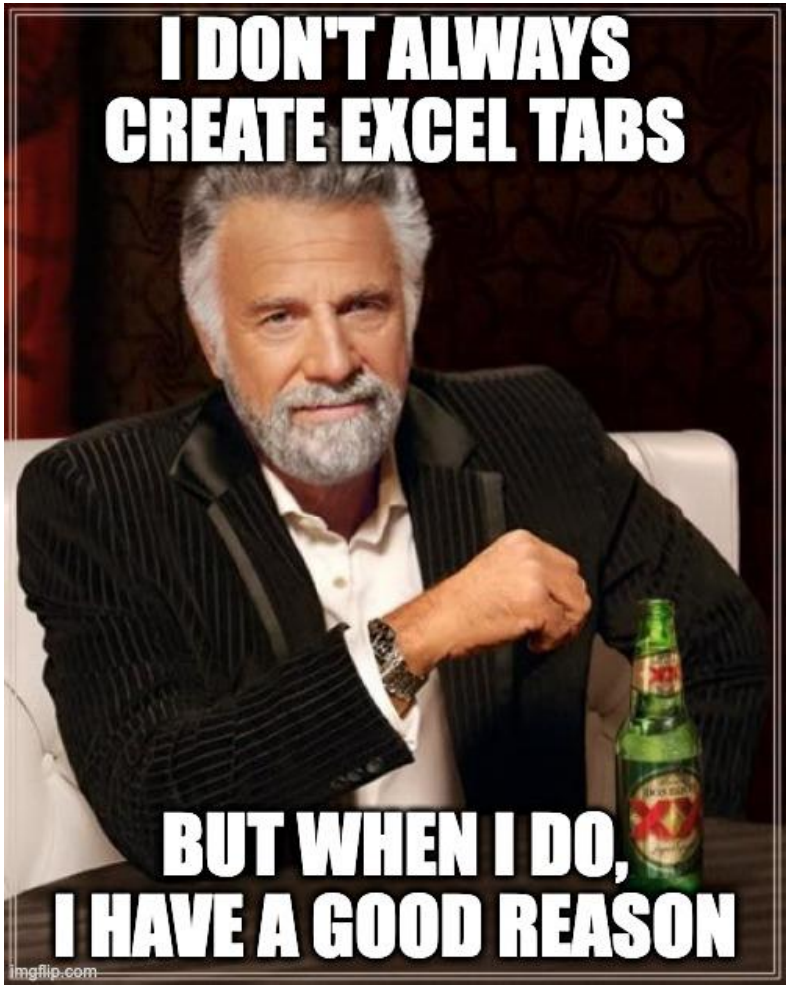


Restrictions with .csv files

- **One csv file per tab**: if your data is in multiple tabs you have to export them separately
  - Creates extra work

Use tabs wisely

- Keep the **raw data** raw
  - Allows you to go back if you make a mistake
- Create a **processed data** tab
  - Keeps all the data in one exportable place
- Create a **notes** tab
  - keep a log of how you processed the raw data with the data itself



How does a computer read...?

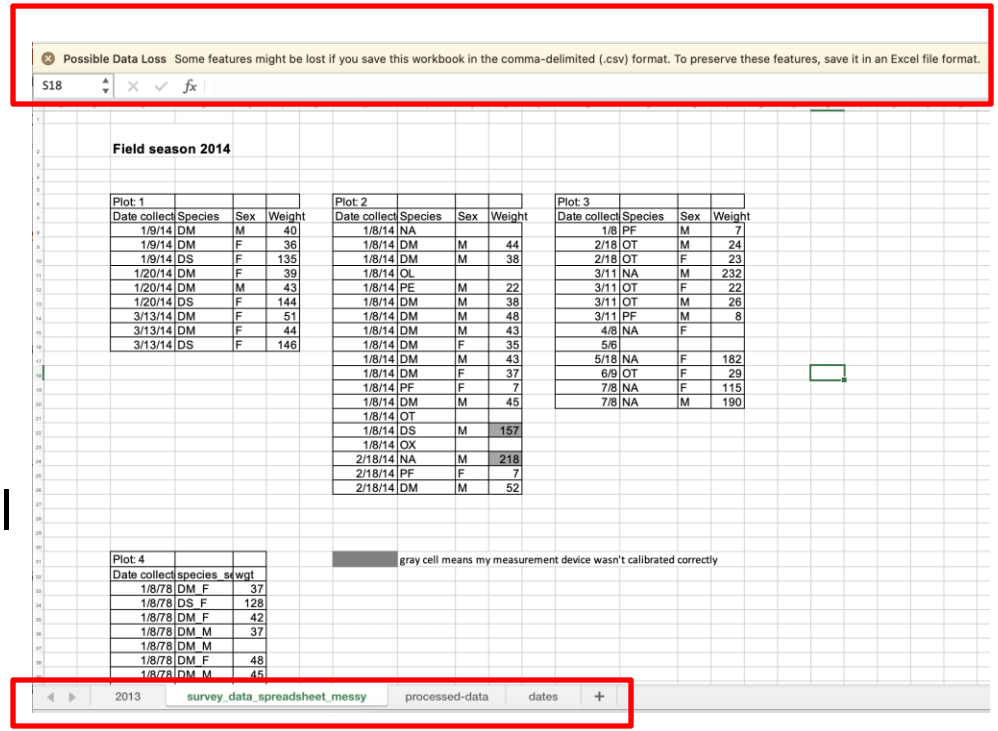
- **Multiple tabs**: it doesn't, keep one tab for processed data, notes and raw data in extra tabs
- **Formatting:**
- **Multiple tables**:
- **Columns**:
- **Missing data**:

Demo: Save to .csv

Result:

- Data Loss warning
- Tabs and formatting are still visible

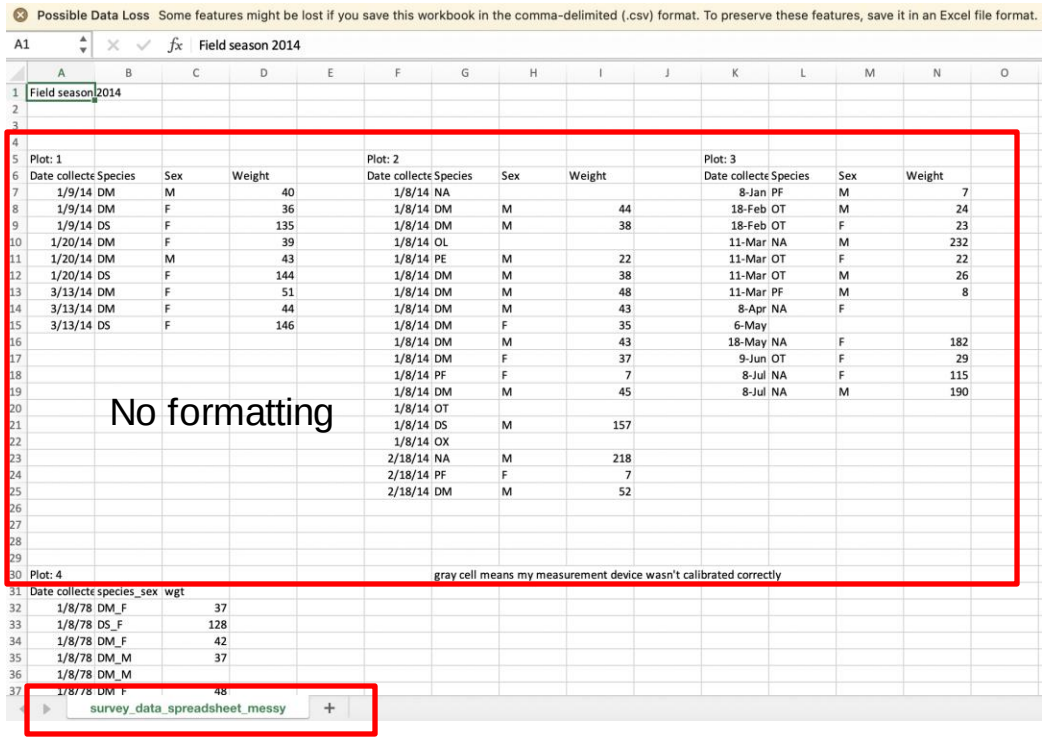
Warning Message



Tabs still present

Demo: Save to .csv

- Open the 2014 tab
- File > Save as
- File Format: csv
- Close the file
- Re-open .csv in spreadsheet program



One Tab

Data Loss

Spreadsheet

Plot: 2			
Date collect	Species	Sex	Weight
1/8/14	NA		
1/8/14	DM	M	44
1/8/14	DM	M	38
1/8/14	OL		
1/8/14	PE	M	22
1/8/14	DM	M	38
1/8/14	DM	M	48
1/8/14	DM	M	43
1/8/14	DM	F	35
1/8/14	DM	M	43
1/8/14	DM	F	37
1/8/14	PF	F	7
1/8/14	DM	M	45
1/8/14	OT		
1/8/14	DS	M	157
1/8/14	OX		
2/18/14	NA	M	218
2/18/14	PF	F	7
2/18/14	DM	M	52

Plot: 3			
Date collect	Species	Sex	Weight
1/8	PF	M	7
2/18	OT	M	24
2/18	OT	F	23
3/11	NA	M	232
3/11	OT	F	22
3/11	OT	M	26
3/11	PF	M	8
4/8	NA	F	
5/6			
5/18	NA	F	182
6/9	OT	F	29
7/8	NA	F	115
7/8	NA	M	190

gray cell means my measurement device wasn't calibrated correctly

.csv

Plot: 2				Plot: 3			
Date collecte	Species	Sex	Weight	Date collecte	Species	Sex	Weight
1/8/14	NA			8-Jan	PF	M	7
1/8/14	DM	M	44	18-Feb	OT	M	24
1/8/14	DM	M	38	18-Feb	OT	F	23
1/8/14	OL			11-Mar	NA	M	232
1/8/14	PE	M	22	11-Mar	OT	F	22
1/8/14	DM	M	38	11-Mar	OT	M	26
1/8/14	DM	M	48	11-Mar	PF	M	8
1/8/14	DM	M	43	8-Apr	NA	F	
1/8/14	DM	F	35	6-May			
1/8/14	DM	M	43	18-May	NA	F	182
1/8/14	DM	F	37	9-Jun	OT	F	29
1/8/14	PF	F	7	8-Jul	NA	F	115
1/8/14	DM	M	45	8-Jul	NA	M	190
1/8/14	OT						
1/8/14	DS	M	157				
1/8/14	OX						
2/18/14	NA	M	218				
2/18/14	PF	F	7				
2/18/14	DM	M	52				

gray cell means my measurement device wasn't calibrated correctly

Restrictions with .csv files

- **One csv file per tab**: if your data is in multiple tabs you have to export them separately
  - Creates extra work

- **Removes all the formatting**: saves only the cell values, separated by commas in a text files, one line per row
  - Any calculations, highlighting, graphs, borders, bolding, color coding in your spreadsheet will disappear
  - Can cause **data loss**

Q: How can I make it machine-readable?

Plot: 2			
Date collect	Species	Sex	Weight
1/8/14	NA		
1/8/14	DM	M	44
1/8/14	DM	M	38
1/8/14	OL		
1/8/14	PE	M	22
1/8/14	DM	M	38
1/8/14	DM	M	48
1/8/14	DM	M	43
1/8/14	DM	F	35
1/8/14	DM	M	43
1/8/14	DM	F	37
1/8/14	PF	F	7
1/8/14	DM	M	45
1/8/14	OT		
1/8/14	DS	M	157
1/8/14	OX		
2/18/14	NA	M	218
2/18/14	PF	F	7
2/18/14	DM	M	52

Plot: 3			
Date collect	Species	Sex	Weight
1/8	PF	M	7
2/18	OT	M	24
2/18	OT	F	23
3/11	NA	M	232
3/11	OT	F	22
3/11	OT	M	26
3/11	PF	M	8
4/8	NA	F	
5/6			
5/18	NA	F	182
6/9	OT	F	29
7/8	NA	F	115
7/8	NA	M	190

gray cell means my measurement device wasn't calibrated correctly

Plot: 2			
Date collecte	Species	Sex	Weight
1/8/14	NA		
1/8/14	DM	M	44
1/8/14	DM	M	38
1/8/14	OL		
1/8/14	PE	M	22
1/8/14	DM	M	38
1/8/14	DM	M	48
1/8/14	DM	M	43
1/8/14	DM	F	35
1/8/14	DM	M	43
1/8/14	DM	F	37
1/8/14	PF	F	7
1/8/14	DM	M	45
1/8/14	OT		
1/8/14	DS	M	157
1/8/14	OX		
2/18/14	NA	M	218
2/18/14	PF	F	7
2/18/14	DM	M	52

Plot: 3			
Date collecte	Species	Sex	Weight
8-Jan	PF	M	7
18-Feb	OT	M	24
18-Feb	OT	F	23
11-Mar	NA	M	232
11-Mar	OT	F	22
11-Mar	OT	M	26
11-Mar	PF	M	8
8-Apr	NA	F	
6-May			
18-May	NA	F	182
9-Jun	OT	F	29
8-Jul	NA	F	115
8-Jul	NA	M	190

gray cell means my measurement device wasn't calibrated correctly

A: Add a column

Plot: 2			
Date collect	Species	Sex	Weight
1/8/14	NA		
1/8/14	DM	M	44
1/8/14	DM	M	38
1/8/14	OL		
1/8/14	PE	M	22
1/8/14	DM	M	38
1/8/14	DM	M	48
1/8/14	DM	M	43
1/8/14	DM	F	35
1/8/14	DM	M	43
1/8/14	DM	F	37
1/8/14	PF	F	7
1/8/14	DM	M	45
1/8/14	OT		
1/8/14	DS	M	157
1/8/14	OX		
2/18/14	NA	M	218
2/18/14	PF	F	7
2/18/14	DM	M	52

Plot: 3			
Date collect	Species	Sex	Weight
1/8	PF	M	7
2/18	OT	M	24
2/18	OT	F	23
3/11	NA	M	232
3/11	OT	F	22
3/11	OT	M	26
3/11	PF	M	8
4/8	NA	F	
5/6			
5/18	NA	F	182
6/9	OT	F	29
7/8	NA	F	115
7/8	NA	M	190

gray cell means my measurement device wasn't calibrated correctly

Date collecte	Species	Sex	Weight	Calibrated
1/8/14	NA			
1/8/14	DM	M	44	Y
1/8/14	DM	M	38	Y
1/8/14	OL			
1/8/14	PE	M	22	Y
1/8/14	DM	M	38	Y
1/8/14	DM	M	48	Y
1/8/14	DM	M	43	Y
1/8/14	DM	F	35	Y
1/8/14	DM	M	43	Y
1/8/14	DM	F	37	Y
1/8/14	PF	F	7	Y
1/8/14	DM	M	45	Y
1/8/14	OT			
1/8/14	DS	M	157	N
1/8/14	OX			
2/18/14	NA	M	218	N
2/18/14	PF	F	7	Y
2/18/14	DM	M	52	Y

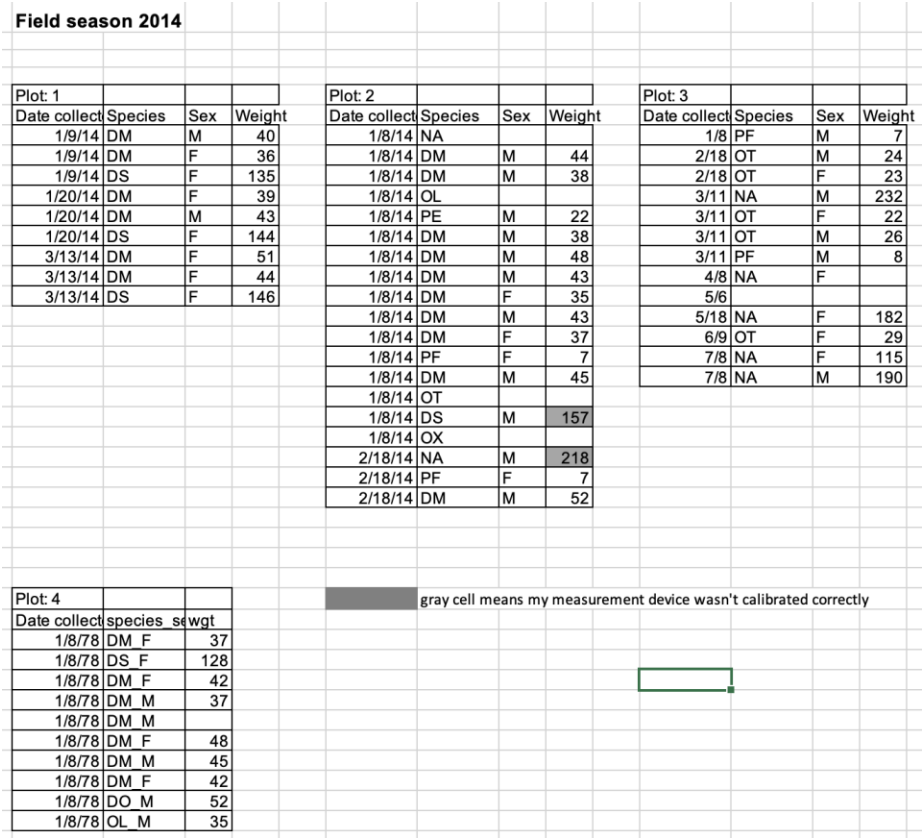
How does a computer read...?

- **Multiple tabs**: it doesn't, keep one tab for processed data, notes and raw data in extra tabs
- **Formatting:** it doesn't, add a column
- **Multiple tables**:
- **Columns**:
- **Missing data**:

Multiple tables

Computers expect

- One **rectangular** data table
- One **value** per cell
- One **observation** per row
- One **variables** per column



Field season 2014

Plot: 1

Date collect	Species	Sex	Weight
1/9/14	DM	M	40
1/9/14	DM	F	36
1/9/14	DS	F	135
1/20/14	DM	F	39
1/20/14	DM	M	43
1/20/14	DS	F	144
3/13/14	DM	F	51
3/13/14	DM	F	44
3/13/14	DS	F	146

Plot: 2

Date collect	Species	Sex	Weight
1/8/14	NA		
1/8/14	DM	M	44
1/8/14	DM	M	38
1/8/14	OL		
1/8/14	PE	M	22
1/8/14	DM	M	38
1/8/14	DM	M	48
1/8/14	DM	M	43
1/8/14	DM	F	35
1/8/14	DM	M	43
1/8/14	DM	F	37
1/8/14	PF	F	7
1/8/14	DM	M	45
1/8/14	OT		
1/8/14	DS	M	157
1/8/14	OX		
2/18/14	NA	M	218
2/18/14	PF	F	7
2/18/14	DM	M	52

Plot: 3

Date collect	Species	Sex	Weight
1/8	PF	M	7
2/18	OT	M	24
2/18	OT	F	23
3/11	NA	M	232
3/11	OT	F	22
3/11	OT	M	26
3/11	PF	M	8
4/8	NA	F	
5/6			
5/18	NA	F	182
6/9	OT	F	29
7/8	NA	F	115
7/8	NA	M	190

Plot: 4

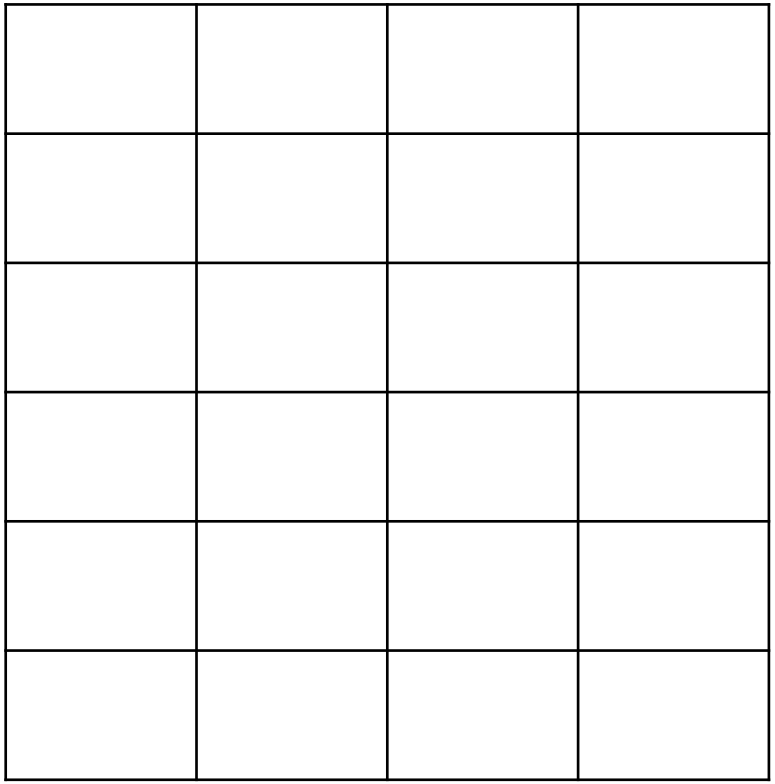
Date collect	species_se	wgt
1/8/78	DM_F	37
1/8/78	DS_F	128
1/8/78	DM_F	42
1/8/78	DM_M	37
1/8/78	DM_M	
1/8/78	DM_F	48
1/8/78	DM_M	45
1/8/78	DM_F	42
1/8/78	DO_M	52
1/8/78	OL_M	35

gray cell means my measurement device wasn't calibrated correctly

Think like a computer

Computers expect

- One **rectangular** data table
- One **value** per cell
- One **observation** per row
- One **variable** per column



Think like a computer

Computers expect

- One **rectangular** data table
- One **value** per cell
- One **observation** per row
- One **variable** per column

AS			
AB	M	23.4	12
AB			
NL	F	23.1	15
AS	M	40.2	13
AB	F	23.5	

Think like a computer

Computers expect

- One **rectangular** data table
- One **value** per cell
- One **observation** per row
- One **variable** per column

Animal1	AS			
Animal2	AB	M	23.4	12
Animal3	AB			
Animal4	NL	F	23.1	15
Animal5	AS	M	40.2	13
Animal6	AB	F	23.5	

Think like a computer

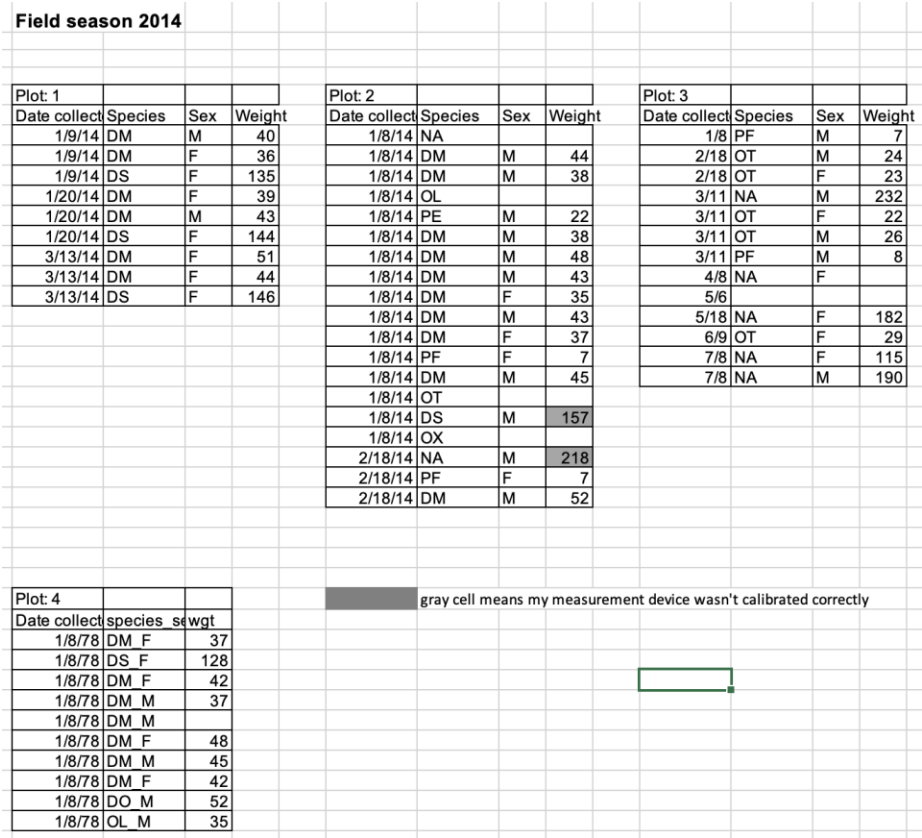
Computers expect

- One **rectangular** data table
- One **value** per cell
- One **observation** per row
- One **variable** per column

	species	sex	weight	length
Animal1	AS			
Animal2	AB	M	23.4	12
Animal3	AB			
Animal4	NL	F	23.1	15
Animal5	AS	M	40.2	13
Animal6	AB	F	23.5	

Multiple tables

- Read as one big table
- Rows contain multiple observation
- Doesn't have all the values for one variable in a column
- Not analyzable



Field season 2014

Plot: 1

Date collect	Species	Sex	Weight
1/9/14	DM	M	40
1/9/14	DM	F	36
1/9/14	DS	F	135
1/20/14	DM	F	39
1/20/14	DM	M	43
1/20/14	DS	F	144
3/13/14	DM	F	51
3/13/14	DM	F	44
3/13/14	DS	F	146

Plot: 2

Date collect	Species	Sex	Weight
1/8/14	NA		
1/8/14	DM	M	44
1/8/14	DM	M	38
1/8/14	OL		
1/8/14	PE	M	22
1/8/14	DM	M	38
1/8/14	DM	M	48
1/8/14	DM	M	43
1/8/14	DM	F	35
1/8/14	DM	M	43
1/8/14	DM	F	37
1/8/14	PF	F	7
1/8/14	DM	M	45
1/8/14	OT		
1/8/14	DS	M	157
1/8/14	OX		
2/18/14	NA	M	218
2/18/14	PF	F	7
2/18/14	DM	M	52

Plot: 3

Date collect	Species	Sex	Weight
1/8	PF	M	7
2/18	OT	M	24
2/18	OT	F	23
3/11	NA	M	232
3/11	OT	F	22
3/11	OT	M	26
3/11	PF	M	8
4/8	NA	F	
5/6			
5/18	NA	F	182
6/9	OT	F	29
7/8	NA	F	115
7/8	NA	M	190

Plot: 4

Date collect	species_se	wgt
1/8/78	DM_F	37
1/8/78	DS_F	128
1/8/78	DM_F	42
1/8/78	DM_M	37
1/8/78	DM_M	
1/8/78	DM_F	48
1/8/78	DM_M	45
1/8/78	DM_F	42
1/8/78	DO_M	52
1/8/78	OL_M	35

gray cell means my measurement device wasn't calibrated correctly

Q: What should the table look like?

Field season 2014

Plot: 1			
Date collect	Species	Sex	Weight
1/9/14	DM	M	40
1/9/14	DM	F	36
1/9/14	DS	F	135
1/20/14	DM	F	39
1/20/14	DM	M	43
1/20/14	DS	F	144
3/13/14	DM	F	51
3/13/14	DM	F	44
3/13/14	DS	F	146

Plot: 2			
Date collect	Species	Sex	Weight
1/8/14	NA		
1/8/14	DM	M	44
1/8/14	DM	M	38
1/8/14	OL		
1/8/14	PE	M	22
1/8/14	DM	M	38
1/8/14	DM	M	48
1/8/14	DM	M	43
1/8/14	DM	F	35
1/8/14	DM	M	43
1/8/14	DM	F	37
1/8/14	PF	F	7
1/8/14	DM	M	45
1/8/14	OT		
1/8/14	DS	M	157
1/8/14	OX		
2/18/14	NA	M	218
2/18/14	PF	F	7
2/18/14	DM	M	52

Plot: 3			
Date collect	Species	Sex	Weight
1/8	PF	M	7
2/18	OT	M	24
2/18	OT	F	23
3/11	NA	M	232
3/11	OT	F	22
3/11	OT	M	26
3/11	PF	M	8
4/8	NA	F	
5/6			
5/18	NA	F	182
6/9	OT	F	29
7/8	NA	F	115
7/8	NA	M	190

- **How many animals were surveyed?** -> # rows
- **What values are being collected for each animal?** -> Variables

Plot: 4		
Date collect	species_se	wgt
1/8/78	DM_F	37
1/8/78	DS_F	128
1/8/78	DM_F	42
1/8/78	DM_M	37
1/8/78	DM_M	
1/8/78	DM_F	48
1/8/78	DM_M	45
1/8/78	DM_F	42
1/8/78	DO_M	52
1/8/78	OL_M	35

gray cell means my measuremer

2013 Field Season

Species: DM			
Date Collecte	Plot	Sex	Weight
7/16/13	2	F	
7/16/13	7	M	33g
7/16/13	3	M	
7/16/13	1	M	
7/18/13	3	M	40g
7/18/13	7	M	48g
7/18/13	4	F	29g
7/18/13	4	F	46g
7/18/13	7	M	36g
7/18/13	7	F	35g
7/18/13	8	F	22g
7/18/13	7	F	42g
7/18/13	4	F	41g
7/18/13	6	F	37g

Species: DO			
Date Collecte	Plot	Sex	Weight
8/19/13	8	F	52
10/17/13	3	F	33
10/17/13	3	F	50
10/17/13	17	F	48
10/17/13	17	F	31
10/18/13	8	F	41
11/12/13	1	F	44
11/12/13	1	M	48
11/14/13	8	F	39
12/10/13	9	F	40
12/10/13	1	M	45
12/11/13	8	F	41

Species: DS			
Date Collecte	Plot	Sex	Weight
11/12/13	9	F	117
11/12/13	1	F	121
11/12/13	20	M	115
11/12/13	9	F	120
11/13/13	17	F	118
11/13/13	11	F	126
11/13/13	17	M	132 (scale not calibrated)
11/13/13	14	F	113 (scale not callibtrated)
11/13/13	11	F	122
11/13/13	4	F	107
11/13/13	4	F	115

Breakout: Create a table

- Make a new tab called "processed-data"
- Combine the data from the 2013 and 2014 tabs into one table
  - **How many animals were surveyed?** -> # rows
  - **What values are being collected for each animal?** -> Variables
- What questions came up for you while you were combining the data?

A: 88 Rows, 6 Variables

- **78 animals surveyed**
- Data **collected for each animal:** Date Collected, Plot, Sex, Weight, Species, Calibrated
- Questions?

	Date Collected	Plot	Sex	Weight	Species	Calibrated?
	7/16/13	2	F		DM	
	7/16/13	7	M	33g	DM	
	7/16/13	3	M		DM	
	7/16/13	1	M		DM	
	7/18/13	3	M	40g	DM	
	7/18/13	7	M	48g	DM	
	7/18/13	4	F	29g	DM	
	7/18/13	4	F	46g	DM	
)	7/18/13	7	M	36g	DM	
1	7/18/13	7	F	35g	DM	
2	7/18/13	8	F	22g	DM	
3	7/18/13	7	F	42g	DM	
4	7/18/13	4	F	41g	DM	
5	7/18/13	6	F	37g	DM	
6	8/19/13	8	F	52	DO	
7	10/17/13	3	F	33	DO	
8	10/17/13	3	F	50	DO	
9	10/17/13	17	F	48	DO	
0	10/17/13	17	F	31	DO	
1	10/18/13	8	F	41	DO	
2	11/12/13	1	F	44	DO	
3	11/12/13	1	M	48	DO	
4	11/14/13	8	F	39	DO	
5	12/10/13	9	F	40	DO	
6	12/10/13	1	M	45	DO	
7	12/11/13	8	F	41	DO	
8	11/12/13	9	F	117	DS	

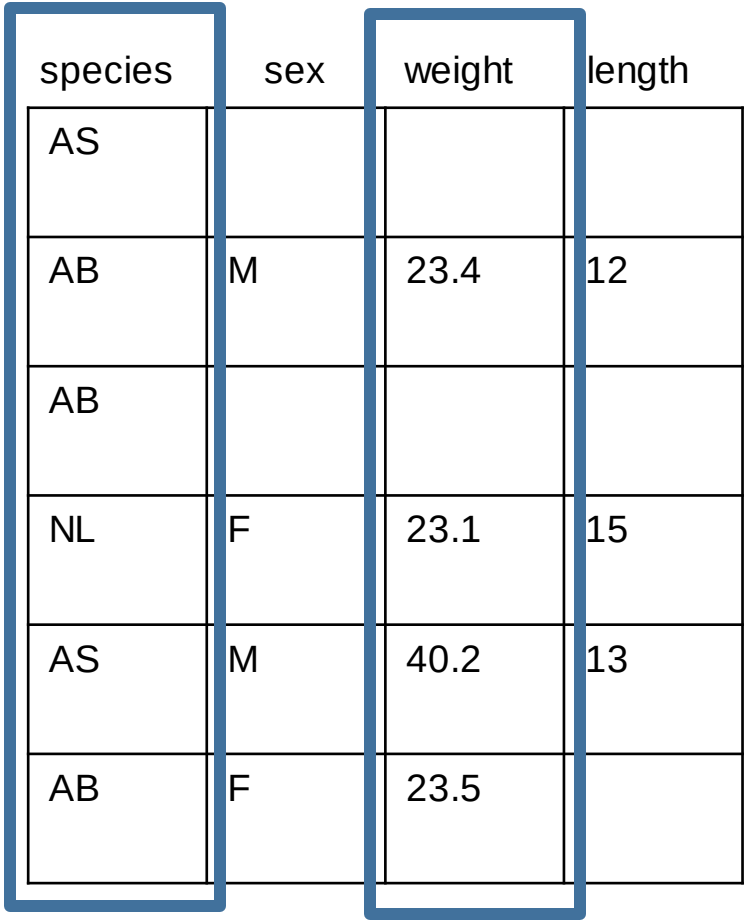
How does a computer read...?

- **Multiple tabs**: it doesn't, keep one tab for processed data, notes and raw data in extra tabs
- **Formatting:** it doesn't, add a column
- **Multiple tables**: as one big table, combine them!
- **Columns**:
- **Missing data**:

Columns

Computers expect

- one **variable** per column
- All values have the same type/format
  - Text = categories
  - Numbers = math



species	sex	weight	length
AS			
AB	M	23.4	12
AB			
NL	F	23.1	15
AS	M	40.2	13
AB	F	23.5	

Column names

- Column headers become variable names
- **Human readable**: Aim for descriptive name
  - Avoid abbreviations
- **Machine readable**: Avoid spaces and most special characters
- Be consistent!

Naming convention	Example
Camel case	speciesName
Snake case	species_name
Kabob case	species-name
Dot case	species.name*

* Can cause issues in Python with pandas

Q: How can we improve this table?

- Is each column a separate variable?
- Is each value in the column the same type
- Can a computer read the column headers?

Date collected	plot	Species-sex	Weight
Jan 9, 1978	1	DM-M	40
1/9/1978	1	DM-F	36 g
1/9/78	1	DS-F	135
1/20/78	2	DM-M	38g
1/20/78	2	DS-f	.144 kg
03/13/1978	2	DM-F	44
3/13/78	2	DS-F	146

A: Structural Changes

- Use consistent header format
- Make a column for each variable

date_collected	plot	species	sex	weight
Jan 9, 1978	1	DM	M	40
1/9/1978	1	DM	F	36 g
1/9/78	1	DS	F	135
1/20/78	2	DM	M	38g
1/20/78	2	DS	f	.144 kg
03/13/1978	2	DM	F	44
3/13/78	2	DS	F	146

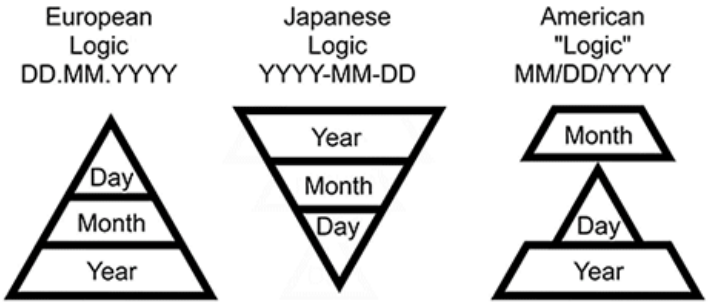
A: Content changes

- Make sure codes for male/female are consistent
- Use a consistent format within columns
- Make sure weights are numbers

date_collected	plot	species	sex	weight_g
1/9/1978	1	DM	M	40
1/9/1978	1	DM	F	36
1/9/1978	1	DS	F	135
1/20/1978	2	DM	M	38
1/20/1978	2	DS	**F**	144
3/13/1978	2	DM	F	44
3/13/1978	2	DS	F	146

Dates are complicated

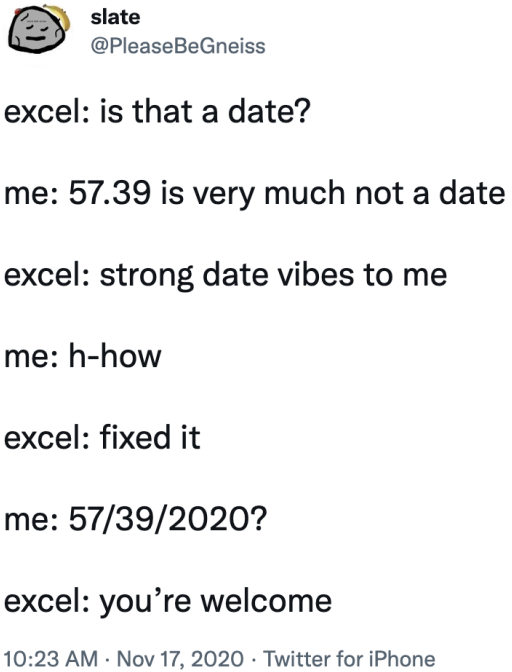
- Date format varies by region
- YYYY vs YY
- Some include timestamps



Dates in Spreadsheets

Spreadsheet software...

- Interprets everything possible as a date
- Date systems vary between versions of same software
- Make assumptions



https://twitter.com/pleasebegneiss/status/1328735477923336192

Example: gene names

- Almost 1/3 of genomics papers have errors in gene names related to date "autocorrecting"
- The gene names that resemble dates (MARCH-1, SEPT-1) have been changed to prevent this from happening

MICROSOFT | REPORT | SCIENCE

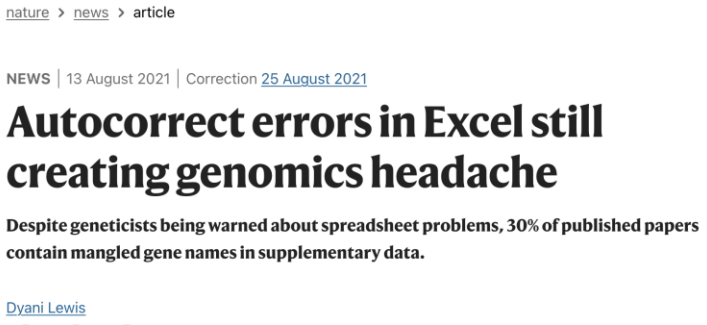
99

Scientists rename human genes to stop Microsoft Excel from misreading them as dates

Sometimes it's easier to rewrite genetics than update Excel

By James Vincent | Aug 6, 2020, 8:44am EDT

https://www.theverge.com/2020/8/6/21355674/human-genes-rename-microsoft-excel-misreading-dates



nature › news › article

NEWS | 13 August 2021 | Correction 25 August 2021

Autocorrect errors in Excel still creating genomics headache

Despite geneticists being warned about spreadsheet problems, 30% of published papers contain mangled gene names in supplementary data.

Dyani Lewis

https://www.nature.com/articles/d41586-021-02211-4

Demo: What year is it?

- Open the "dates" tab in survey_data_spreadsheet_messy.xls
  - What year were the measurements taken?

- Click File> Save As > Select .csv UTF-8 from the File Format Dropdown

- Open the .csv file in a text editor (TextEdit or Notepad)
  - What year were the measurements taken?

- Open the .csv file in a spreadsheet program (Excel, Numbers, Google Sheet)
  - What year were the measurements taken?

A: Safe dates

- Separate date into Day – Month – Year columns
- Can be recombined later

year	month	day	plot	species	sex	weight_g
1978	1	9	1	DM	M	40
1978	1	9	1	DM	F	36
1978	1	9	1	DS	F	135
1978	1	20	2	DM	M	38
1978	1	20	2	DS	F	144
1978	3	13	2	DM	F	44
1978	3	13	2	DS	F	146

How does a computer read...?

- **Multiple tabs**: it doesn't, keep one tab for processed data, notes and raw data in extra tabs
- **Formatting:** it doesn't, add a column
- **Multiple tables**: as one big table,
- **Columns**: one variable per column, use consistent formatting
  - Be careful with dates
- **Missing data**:

Missing Data

- **Example**: A surveyed animal escapes after you identify the species and sex but before you can weigh it

- How do you record this?
  - Use a Null Value

https://www.redlandsdailyfacts.com/2020/08/07/endangered-kangaroo-rats-now-wearing-backpacks-in-wash-near-redlands-highland/

Tips for picking null values

Null values are "symbols" that represent missing data

- Make sure it's not a valid value
- Programming languages have default Null values
  - R: NA
  - SQL: Null
  - Python: None
- **Avoid** numbers (0, 999, -999), special symbols (*,+,-) and uncommon text labels

Common Null values

Value	Compatible Language	Pros	Cons
Blank	tidyverse, pandas, SQL	Unlikely to be a valid value, easily read by common data science languages	Hard to know if it's missing data or accidentally deleted
NA	R, tidyverse, pandas	easily read by common data science languages	Could be a valid value
NULL	SQL	Default for SQL	Could be a valid value
None	Python	Default for Python	Could be a valid value
NaN	Pandas	Default for Pandas	Could be a valid value
Missing	Julia	Default for Julia	Could be a valid value

How does a computer read...?

- **Multiple tabs**: it doesn't, keep one tab for processed data, notes and raw data in extra tabs
- **Formatting:** it doesn't, add a column
- **Multiple tables**: as one big table,
- **Columns**: one variable per column, consistent formatting
  - Be careful with dates
- **Missing data**: it depends! Think about your data structure and what you will use to analyze the data

Need help?

- Email: tobin.magle@northwestern.edu
- Request a consultation: https://app.smartsheet.com/b/form/2f2ec327e6164f83b588b7bbe2e2b56f
- Source material for this lesson: http://www.datacarpentry.org/spreadsheet-ecology-lesson/

Exercise:

- Think about your own spreadsheet data

OR

- Pick an example from the next 2 slides

- How can you make the data machine readable?

Example: Supplemental_data_1_xls

- https://figshare.com/articles/Supplemental_data_1_xls/4055544

- **Description**: “Table of the results given by HPLC analysis of the samples. Key: Rt, retention time; +, presence of peak; -, absence of peak.”

Example: cck8_xls

- https://figshare.com/articles/cck8_xls/3505772

- **Description**: “This data are from CCK-8 assay and ELISA.”