

Lab 1: Tidy data

Learning objectives

- Understand value of organized data
- Observe, validate, and correct datasets

Skills to master

- Load and export Excel, text, csv files
- Clean data
- Organize data into long or wide format.
- Conduct basic numerical transformations and summary statistics

Overview

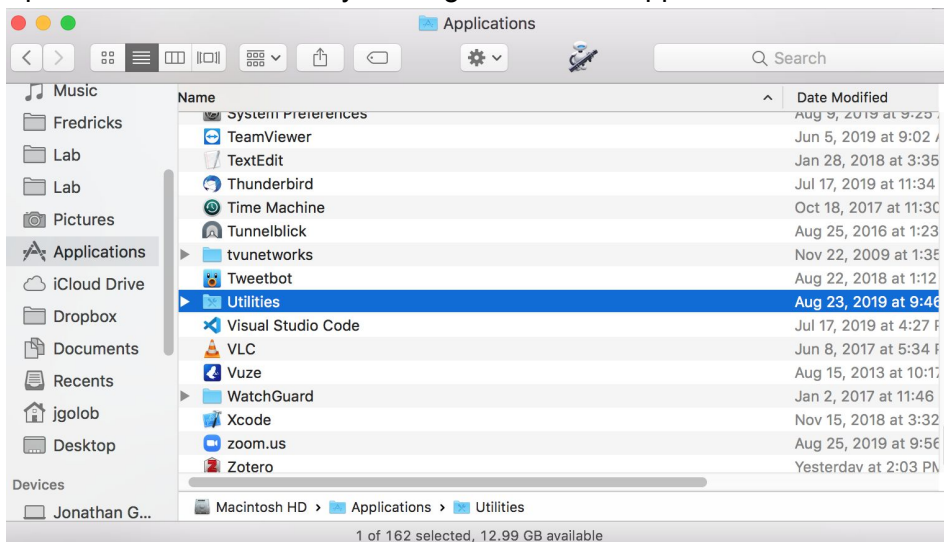
What are clean data?

We generate data to answer questions. How the data are organized can help (or limit) our ability to use the data. Clean data are organized in a way that makes it easy to verify the data are correct, check that there are no bits of information missing, and allow easy to answer questions with the data.

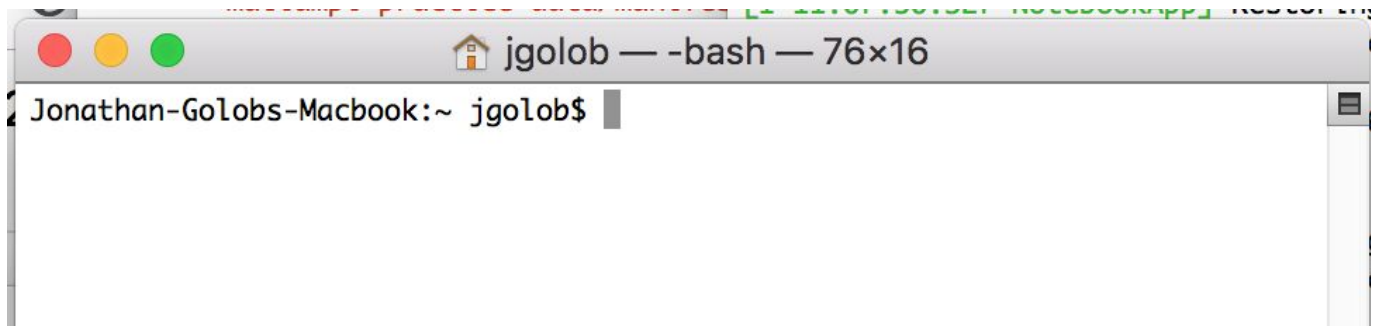
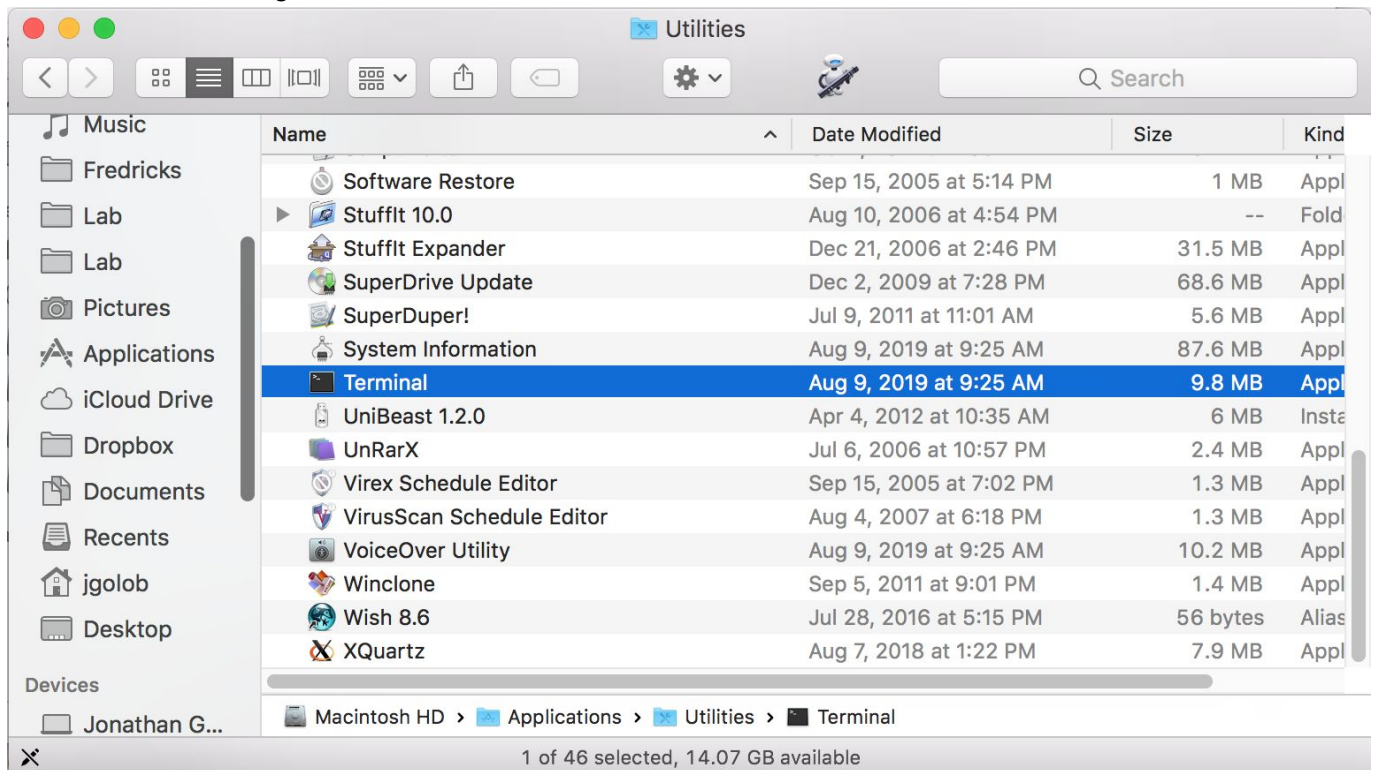
Today we will learn the attributes of clean data as well as tactics to convert raw data into clean data.

I. Keeping track of data with git

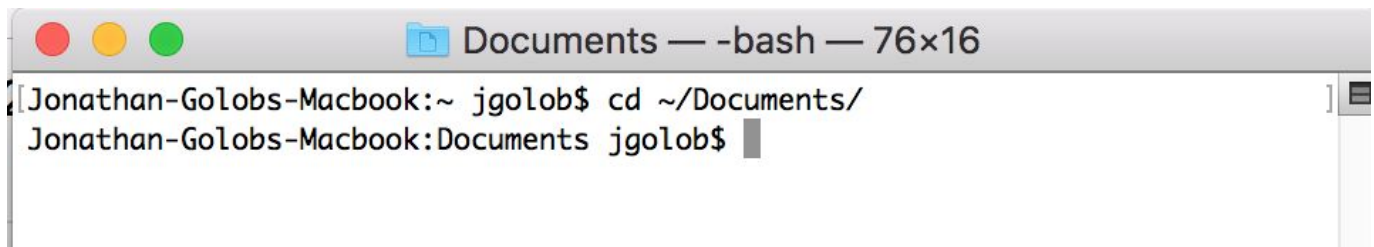
1. If you did not before lab today, make a GitHub account: <https://github.com/> and star the following repository: [gdanetzk/UMich_Bio201_F19/](https://github.com/gdanetzk/UMich_Bio201_F19/)
2. You will have to do this at the beginning of each lab session to download the material for that week.
3. Open a terminal window by clicking on Finder > Applications > Utilities



4. Then selecting Terminal:

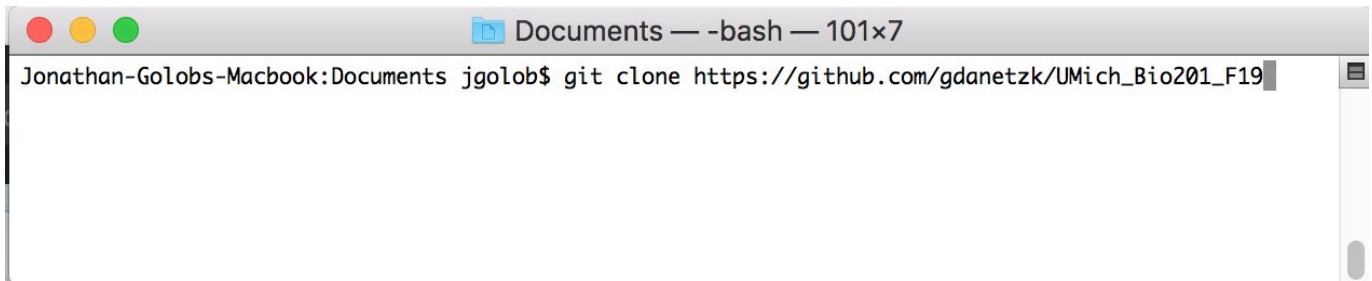


5. Change to your documents directory by typing, "cd ~/Documents"



6. Use git to download the course materials: type in

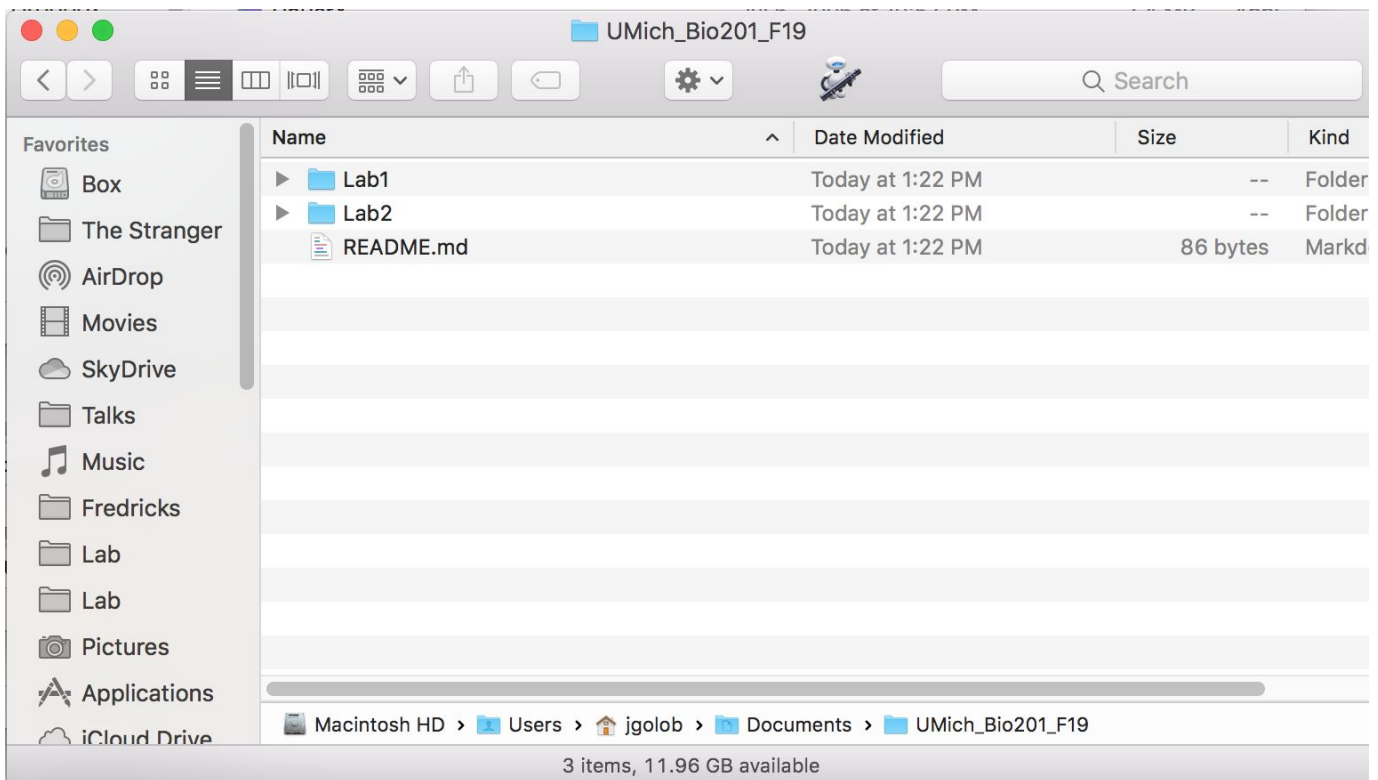
```
git clone https://github.com/gdanetzk/UMich\_Bio201\_F19
```



A terminal window titled "Documents — -bash — 101x7" shows the command `git clone https://github.com/gdanetzk/UMich_Bio201_F19` being entered at the prompt `Jonathan-Golobs-Macbook:Documents jgolob$`.

We are using a program git which can help us download data, keep track of changes, allow us to undo changes or look at older versions of files easily.

7. Type “open .” to open a finder window here, and look at the Bio201 course directory.



The Lab1 folder will contain the data we will use for the remaining protocols.

II. Organizing raw study data into Wide Format

Tasks:

- Count number of individuals
- Count number of samples per individual
- Calculate average weight by individual
- Determine range of weights for cohort

The study:

Six participants were asked to record their weight in kilograms weekly for a month and a half.

The questions:

- What was the average weight of each person during the study?
- What was the range of weights observed during this study?
- How many participants were able to collect weekly weights for at least 30 days?
- What was the average weight of each person during the first 15 days of the study?
- What was the average weight of each person during days 16-30 of the study?
- What was the change in weight from the first 15 day average to the last 15 day average?

Protocol:

1. Open the data document (rrw.docx) in Microsoft Word:

79.5 01-02-2018
88.1|01/03/2018
74.6|20180106
101.7,20180106
95.5,01/06/2018
82.4/01/09/2018
90.9,20180110
97.5:01/11/2018
98.4,20180113
71.6/20180115
79.8|01-16-2018
100.2|01-16-2018

Each participant was assigned a color. In each row of the Word table is the weight and date the weight was collected. This is an anonymized example of an actual un-tidy dataset used by researchers. This is the worst-case scenario for data organization.

Can you think of a strategy to answer any of our questions with these data in this format?

2. Open an empty spreadsheet in Excel.
3. We will turn these data into a spreadsheet, also known as 'wide' format data. In wide format, each specimen / sample / participant gets a row.
4. Assign a study ID to each participant.

Best practices for designing study IDs:

 - i. Pick a prefix that is unique to this study. (e.g. "w60").
 - ii. Use numbers in a sequence that are zero padded (e.g. "w60__001")

The screenshot shows a Microsoft Word document on the left with a table containing 12 rows of data. Each row has a colored background and contains a weight and a date in various formats. On the right, an Excel spreadsheet is shown with a table that has 7 rows and 3 columns: 'color', 'study_id', and an empty column. The data from the Word document is being mapped to the Excel spreadsheet, where each row represents a participant's data.

	A	B	C
1	color	study_id	
2	blue	w60_001	
3	violet	w60_002	
4	red	w60_003	
5	yellow	w60_004	
6	green	w60_005	
7			

- iii. Be consistent in using capitals, and how you separate the prefix from the number
- iv. For studies involving people, do not use anything personally identifying about a participant, such as names, initials, social security numbers, email, etc.
- v. Store somewhere else the answer key linking the study ID to the identifying information for the person (in trials involving people).

Note the column headers. The first row is usually left to describe what the column holds, in this case the color assigned to the participant, and the study ID created. When picking column names (and IDs), it's best to use only one case (lower case here) and have no spaces in the name.

5. Fill in our column names. In this spreadsheet we will make a column (e.g., week_01, week_02, etc.) for each week's measurement of weight for that person.

	A	B	C	D	E	F	G	H	I	J
1	color	study_id	week_01	week_02	week_03	week_04	week_05	week_06	week_07	week_08
2	blue	w60_001								
3	violet	w60_002								
4	red	w60_003								
5	yellow	w60_004								
6	green	w60_005								

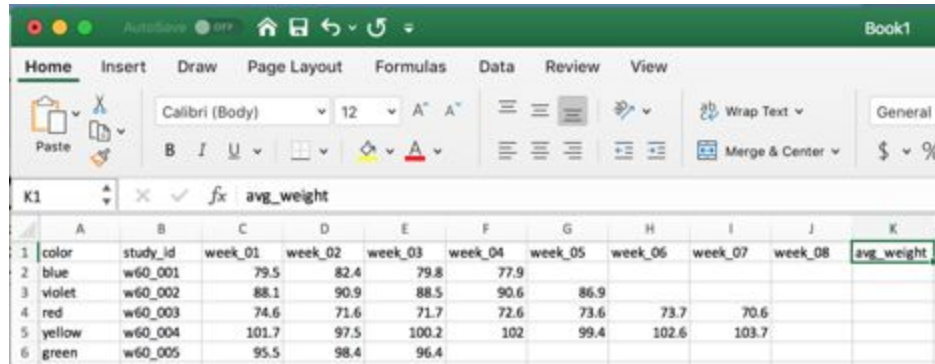
In spreadsheets, each column stores some information about each sample / participant. In this case we are going to store a given study-week weight in additional columns. We already have our color and study_id columns.

6. Now fill in the weights for each person and week. For now do *not* copy over the dates, just the weight.

	A	B	C	D	E	F	G	H	I	J
1	color	study_id	week_01	week_02	week_03	week_04	week_05	week_06	week_07	week_08
2	blue	w60_001	79.5	82.4	79.8	77.9				
3	violet	w60_002	88.1	90.9	88.5	90.6	86.9			
4	red	w60_003	74.6	71.6	71.7	72.6	73.6	73.7	70.6	
5	yellow	w60_004	101.7	97.5	100.2	102	99.4	102.6	103.7	
6	green	w60_005	95.5	98.4	96.4					

- Discuss with your neighbor: What would you do if some of the weights were in pounds and some in kilograms? How would you know? What if some had units and some did not?
- This raw data was at least sorted by date. What if it wasn't? Did you verify each weight was about a week / seven days after the prior? How did you deal with the varying date formats?

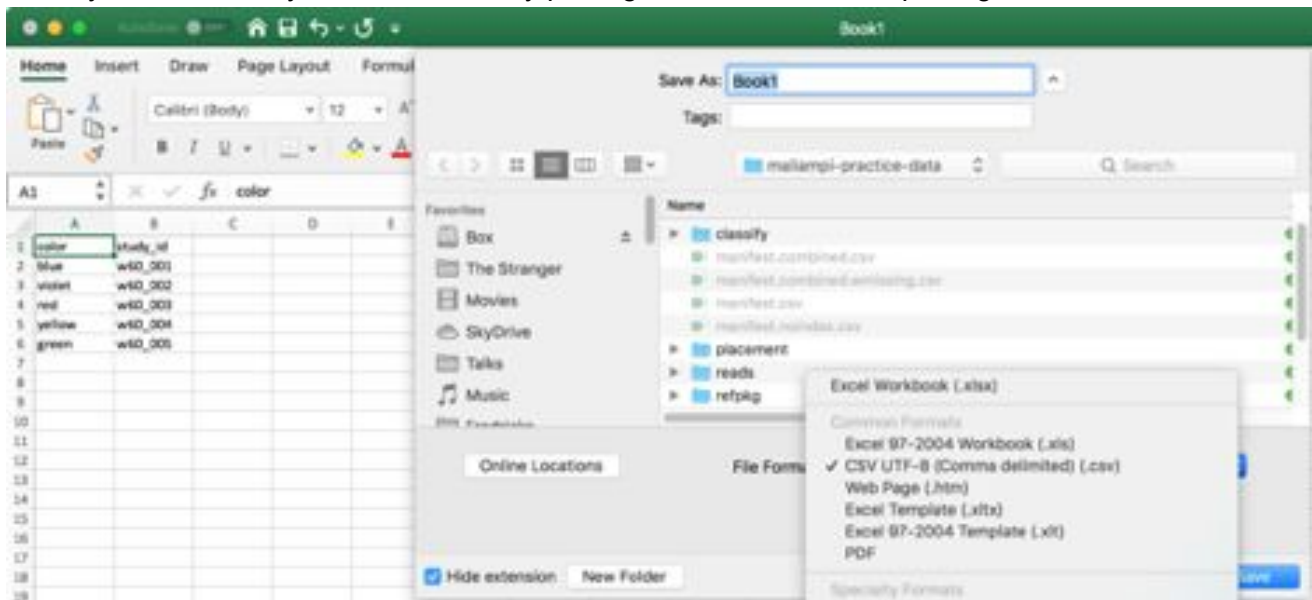
7. The data is now in wide format. Compare to the original document. It's a lot easier to see how each person's weight changed over time. Let's calculate an average weight for each person by adding a column to the wide data that is the average of weights for each person.



	A	B	C	D	E	F	G	H	I	J	K
	color	study_id	week_01	week_02	week_03	week_04	week_05	week_06	week_07	week_08	avg_weight
1	blue	w60_001	79.5	82.4	79.8	77.9					
2	violet	w60_002	88.1	90.9	88.5	90.6	86.9				
3	red	w60_003	74.6	71.6	71.7	72.6	73.6	73.7	70.6		
4	yellow	w60_004	101.7	97.5	100.2	102	99.4	102.6	103.7		
5	green	w60_005	95.5	98.4	96.4						

In wide format, some columns can come directly from the data, others can be made by transforming the data.

8. Save your answer key in **CSV format** by picking File > Save As and picking the format CSV UTF-8



- Excel will complain you are not using the xlsx format now and every time you save in CSV format. Ignore this. Excel is giving you bad advice.
- Excel workbook format is proprietary, and only will work well in Excel. CSV files are usable in more software, simpler, and preferred for use in science.
- Always save a copy of your data before performing any calculations / manipulations on it.

9. We will use excel formulas to calculate an average weight:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	color	study_id	week_01	week_02	week_03	week_04	week_05	week_06	week_07	week_08	avg_weight			
2	blue	w60_001	79.5	82.4	79.8	77.9					=average(C2:J2)			
3	violet	w60_002	88.1	90.9	88.5	90.6	86.9							
4	red	w60_003	74.6	71.6	71.7	72.6	73.6	73.7	70.6					
5	yellow	w60_004	101.7	97.5	100.2	102	99.4	102.6	103.7					
6	green	w60_005	95.5	98.4	96.4									

10. In column K, put in “=AVERAGE(C2:J2)”. Hit enter.

11. Grab the little box in the lower right corner of cell K2 and drag it down to K6.

This applies the same formula to each row.

	J	K
1	week_08	avg_weight
2		79.9

12. Similarly, let's calculate an average weight for week 1 and week 2 and an average weight for week 3 and week 4.

- Make column names for column L and M (avg_wk12 and avg_wk34).
- Put the formula ‘=AVERAGE(C2:D2)’ into cell L2, and ‘=AVERAGE(E2:F2)’ into cell M2.
- Select cells L2 and M2; grab the box in the lower right and drag down.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	color	study_id	week_01	week_02	week_03	week_04	week_05	week_06	week_07	week_08	avg_weight	avg_wk12	avg_wk34
2	blue	w60_001	79.5	82.4	79.8	77.9					79.9	80.95	78.85
3	violet	w60_002	88.1	90.9	88.5	90.6	86.9				89	89.5	89.55
4	red	w60_003	74.6	71.6	71.7	72.6	73.6	73.7	70.6		72.6285714	73.1	72.15
5	yellow	w60_004	101.7	97.5	100.2	102	99.4	102.6	103.7		101.014286	99.6	101.1
6	green	w60_005	95.5	98.4	96.4						96.7666667	96.95	96.4

We now have calculated an average weight for each participant over the first two weeks and for week 3 and 4.

13. Let's now calculate the difference between the week 3 and 4 average to the week 1 and 2 average for each person. We will add a new column name into N1: wk34vswk12.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	color	study_id	week_01	week_02	week_03	week_04	week_05	week_06	week_07	week_08	avg_weight	avg_wk12	avg_wk34	wk34vswk12
2	blue	w60_001	79.5	82.4	79.8	77.9					79.9	80.95	78.85	-2.1
3	violet	w60_002	88.1	90.9	88.5	90.6	86.9				89	89.5	89.55	
4	red	w60_003	74.6	71.6	71.7	72.6	73.6	73.7	70.6		72.6285714	73.1	72.15	
5	yellow	w60_004	101.7	97.5	100.2	102	99.4	102.6	103.7		101.014286	99.6	101.1	
6	green	w60_005	95.5	98.4	96.4						96.7666667	96.95	96.4	

14. We will put the new formula into N2 “=M2 – L2”

15. Just like before, after selecting N2, we can grab the little box in the bottom right and pull it down to N6.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	color	study_id	week_01	week_02	week_03	week_04	week_05	week_06	week_07	week_08	avg_weight	avg_wk12	avg_wk34	wk34-avgwk12
2	blue	w60_001	79.5	82.4	79.8	77.9					79.9	80.95	78.85	-2.1
3	violet	w60_002	88.1	90.9	88.5	90.6	88.9				89	89.5	89.55	0.05
4	red	w60_003	74.6	71.4	71.7	72.6	71.6	71.7	70.6		72.6385714	73.1	72.15	-0.95
5	yellow	w60_004	101.7	97.5	100.2	102	99.4	102.6	103.7		101.014386	99.6	101.1	1.5
6	green	w60_005	95.5	98.4	96.4						96.7666667	96.95	96.4	-0.55

We have now organized our data into “wide format”, sometimes called a spreadsheet. Wide format is great for doing the same thing for each set of data, transforming and normalizing data. Wide format does have some limits, including some we’ve encountered here (what did we do with all the dates!?)

III. Organizing data into Long Format

The study: Five participants were asked to record their weight in kilograms weekly for a month and a half (same as section 2).

The questions:

- What was the average weight of each person during the study?
- What was the range of weights observed during this study?
- How many participants able to collect weekly weights for at least 30 days?
- What was the average weight of each person during the first 15 days on the study?
- What was the average weight of each person during days 16-30 on the study?
- What was the change in weight from the first 15 day average to the last 15 day average?

Protocol:

1. Open the data document in Microsoft Word:

79.5 01-02-2018
88.1|01/03/2018
74.6|20180106
101.7,20180106
95.5,01/06/2018
82.4/01/09/2018
90.9,20180110
97.5:01/11/2018
98.4,20180113
71.6/20180115
79.8|01-16-2018
100.2|01-16-2018

Each participant was assigned a color.

In each row of the Word table is the weight and date the weight was collected.

Each row has multiple bits of information (date and weight)

Only one piece of information should be in a cell. Each category of information (date collected, weight, etc) should be in its own column.

2. Open an empty spreadsheet in Excel.
 - a. We will turn this data into long format.

- b. Create the following columns in the first row: “color”, “study_id”, “date”, “weight_kg”

	A	B	C	D
1	color	study_id	date	weight_kg
2				
3				

In long format, each unit of data (in this case a weight) has its own row. The value of the data gets one column, with all the other columns used to categorize the value (which person, which date, etc.). This allows slicing up data into very nuanced pieces, and is very flexible, if a bit harder to read.

It is useful to document units (i.e., weight is in kilograms), but that should be in the column name, not in each bit of data.

- We will reuse the study IDs we came up with in the first protocol (linking color to study_id).
- Now copy over the data from the word document to our long-format in Excel. Each cell in the word table gets one row in the excel table.

This is painfully tedious and prone to error. One way to avoid this is to have data collected in either wide or long format initially rather than having to convert it over later.

- Notice that the dates are in all different formats. Excel auto-converted some of these but not others. Values in a given column (regardless of long or wide formats) need to be exactly the same format/style in order to properly analyze the data. Hand convert the date formats without slashes to match the other dates.

	A	B	C	D
1	color	study_id	date	weight_kg
2	blue	w60_001	1/2/18	79.5
3	violet	w60_002	1/3/18	88.1
4	red	w60_003	20180106	74.6
5	yellow	w60_004	20180106	101.7
6	green	w60_005	1/6/18	95.5
7	blue	w60_001	1/9/18	82.4
8	violet	w60_002	20180110	90.9
9	yellow	w60_004	1/11/18	97.5
10	green	w60_005	20180113	98.4
11	red	w60_003	20180115	71.6
12	blue	w60_001	1/16/18	79.8
13	yellow	w60_004	1/16/18	100.2
14	violet	w60_002	1/18/18	88.5
15	green	w60_005	1/20/18	96.4
16	red	w60_003	20180122	71.7
17	blue	w60_001	20180122	77.9
18	yellow	w60_004	20180122	102
19	violet	w60_002	1/25/18	90.6

	A	B	C	D
1	color	study_id	date	weight_kg
2	blue	w60_001	1/2/18	79.5
3	violet	w60_002	1/3/18	88.1
4	red	w60_003	1/6/18	74.6
5	yellow	w60_004	1/6/18	101.7
6	green	w60_005	1/6/18	95.5
7	blue	w60_001	1/9/18	82.4
8	violet	w60_002	1/10/18	90.9
9	yellow	w60_004	1/11/18	97.5
10	green	w60_005	1/13/18	98.4
11	red	w60_003	1/15/18	71.6
12	blue	w60_001	1/16/18	79.8
13	yellow	w60_004	1/16/18	100.2
14	violet	w60_002	1/18/18	88.5
15	green	w60_005	1/20/18	96.4
16	red	w60_003	1/22/18	71.7
17	blue	w60_001	1/22/18	77.9
18	yellow	w60_004	1/22/18	102
19	violet	w60_002	1/25/18	90.6

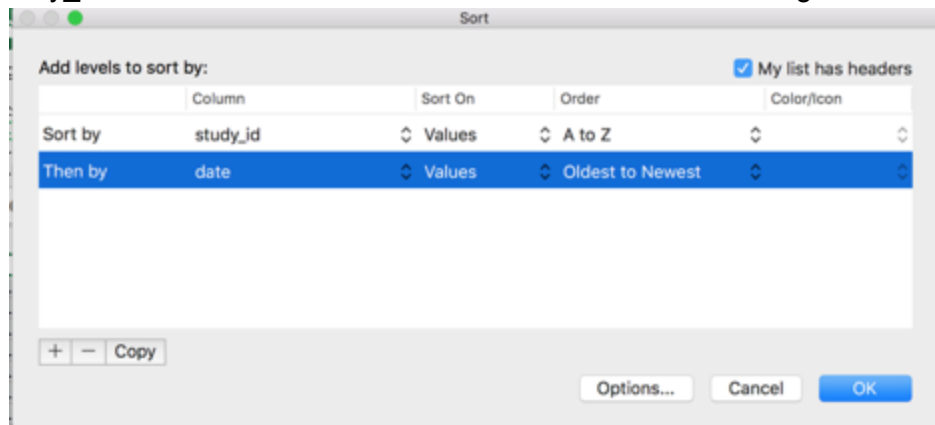
Excel is notorious for auto-converting data, particularly dates, incorrectly (as well as autocorrecting other data, such as gene names). What do you think of it? What problems can you foresee from having data autocorrected by Excel (which is not designed with biology or microbiology in mind)?

6. Save the raw data in CSV format.

7. Sort the rows by study_id then date. Open the “Data” ribbon, then click on “sort”



8. Select the “study_id” column. Then click the “+” icon to add another sorting row. Add “date”



9. click OK and allow the sorting to happen.

	A	B	C	D
1	color	study_id	date	weight_kg
2	blue	w60_001	1/2/18	79.5
3	blue	w60_001	1/9/18	82.4
4	blue	w60_001	1/16/18	79.8
5	blue	w60_001	1/22/18	77.9
6	violet	w60_002	1/3/18	88.1
7	violet	w60_002	1/10/18	90.9
8	violet	w60_002	1/18/18	88.5
9	violet	w60_002	1/25/18	90.6
10	violet	w60_002	2/2/18	86.9
11	red	w60_003	1/6/18	74.6
12	red	w60_003	1/15/18	71.6
13	red	w60_003	1/22/18	71.7
14	red	w60_003	1/31/18	72.6
15	red	w60_003	2/7/18	73.6
16	red	w60_003	2/15/18	73.7
17	red	w60_003	2/23/18	70.6
18	yellow	w60_004	1/6/18	101.7

We answer questions with data in long format by sorting and filtering (later grouping and slicing when we access data with code).

10. Let's add a new column, study_days, which will be how many days into the

	A	B	C	D	E
1	color	study_id	date	weight_kg	study_days
2	blue	w60_001	1/2/18	79.5	
3	blue	w60_001	1/9/18	82.4	
4	blue	w60_001	1/16/18	79.8	
5	blue	w60_001	1/22/18	77.9	
6	violet	w60_002	1/3/18	88.1	
7	violet	w60_002	1/10/18	90.9	

study for that participant the weight was measured.

11. In cell E1, we will put the formula `=C2-C$2`. After hitting enter and selecting the E2 cell, drag the box in the right corner down until we are through all of blue's rows.

	A	B	C	D	E
1	color	study_id	date	weight_kg	study_days
2	blue	w60_001	1/2/18	79.5	0
3	blue	w60_001	1/9/18	82.4	7
4	blue	w60_001	1/16/18	79.8	14
5	blue	w60_001	1/22/18	77.9	20
6	violet	w60_002	1/3/18	88.1	

12. Do the same for violet, red, yellow, green, and teal:

`=C6-C$6` into E6 and then drag down to E10.

`=C11-C$11` into E11 and then drag down to E15.

`=C18-C$18` into E18, dragged through E22

`=C25-C$25` into E25, dragged through E29

`=C28-C$28` into E28, dragged through E32

Excel isn't particularly helpful when looking at the data in this format. When we analyze data with code (such as in R, which we will do later in this course) this format works better.

13. Look over the groups.
- How many participants collected at least 30 days of weights?
 - How many weights were typical from day 0-15 and 16-30?
 - Do you think we would get a more accurate result with data in this format as compared to the wide / spreadsheet format?

	A	B	C	D	E
1	color	study_id	date	weight_kg	study_days
2	blue	w60_001	1/2/18	79.5	0
3	blue	w60_001	1/9/18	82.4	7
4	blue	w60_001	1/16/18	79.8	14
5	blue	w60_001	1/22/18	77.9	20
6	violet	w60_002	1/3/18	88.1	0
7	violet	w60_002	1/10/18	90.9	7
8	violet	w60_002	1/18/18	88.5	15
9	violet	w60_002	1/25/18	90.6	22
10	violet	w60_002	2/2/18	86.9	30
11	red	w60_003	1/6/18	74.6	0
12	red	w60_003	1/15/18	71.6	9
13	red	w60_003	1/22/18	71.7	16
14	red	w60_003	1/31/18	72.6	25
15	red	w60_003	2/7/18	73.6	32
16	red	w60_003	2/15/18	73.7	40
17	red	w60_003	2/23/18	70.6	48
18	yellow	w60_004	1/6/18	101.7	0
19	yellow	w60_004	1/11/18	97.5	5
20	yellow	w60_004	1/16/18	100.2	10
21	yellow	w60_004	1/22/18	102	16
22	yellow	w60_004	1/29/18	99.4	23
23	yellow	w60_004	2/6/18	102.6	31
24	yellow	w60_004	2/13/18	103.7	38
25	green	w60_005	1/6/18	95.5	0
26	green	w60_005	1/13/18	98.4	7
27	green	w60_005	1/20/18	96.4	14
28	teal	w60_006	1/29/18	103.9	0
29	teal	w60_006	2/3/18	106.3	5
30	teal	w60_006	2/12/18	107.5	14
31	teal	w60_006	2/20/18	104.2	22
32	teal	w60_006	3/1/18	106.6	31
33	teal	w60_006	3/8/18	106.6	38

IV. Identify errors in the dataset and correct them

(Adapted from “Data Organization in Spreadsheets” by Karl Broman and Kara Woo

<https://www.tandfonline.com/doi/full/10.1080/00031305.2017.1375989>)

1. Open the dirty_data.csv file in excel.

Possible Data Loss Some features might be lost if you save this workbook in the comma-delimited (.csv) format.

	A	B	C	D	E	F	G	H	I
1					Gas:				
2	Participant ID	Timestamp	Date:	Quintron Bre	Hydrogen:	Methane:	Carbon Dioxi	Correction:	Comments:
3	789	#####	1/27/19	Quintron 2	18	3	3.7	1.48	
4		#####	2/1/19	Quintron 2	26	2	4.8	1.14	
5		#####	2/3/19	Quintron 2	1	1	4.4	1.25	
6		#####	2/3/19	Quintron 2	1	1	4.6	1.19	
7		#####	2/13/19	Quintron 2	72	2	4.5	1.22	
8		#####	2/14/19	Quintron 2	63	2	5.1	1.07	
9		#####	2/17/19	Quintron 2	25	2	4.6	1.19	
10	U700	#####	1/26/19	Quintron 2	0	1	5	1.1	
11		#####	1/29/19	Quintron 1	7	1	3.7	1.48	
12		#####	2/12/19	Quintron 2	13	1	4.6	1.19	
13		#####	2/14/19	Quintron 1	3	0	3.9	1.41	
14	U701	#####	1/29/19	Quintron 2	10	12	4.5	1.22	
15		#####	2/5/19	Quintron 2	45	8	5.3	1.03	
16		#####	2/16/19	Quintron 1	5	5	5.4	1.01	

Note: The “Possible Data Loss” warning is not good advice from Excel.

2. Make the data square.

Notice how the “Participant ID” column isn’t for every row. Instead, only the first entry for each block is filled in. Likewise, the first row only has “Gas:”. Proper wide or long formatted data is ‘square’. There is only one header column and every cell is filled in. For later data analysis it is important to have the participant ID listed in each row, not just the first row for that participant’s data.

- Remove the first row; highlight the whole row, then right click and select 'Delete'.

A1									
	A	B	C	D	E	F	G	H	I
1					Gas:				
2	Participant ID	Timestamp	Date:	Quintron Bre	Hydrogen:	Methane:	Carbon Dioxi	Correction:	Comments:
3	789	#####	1/27/19	Quintron 2	18	3	3.7	1.48	

Select each participant ID and copy it to fill in all the empty rows by clicking on the box in the lower right of the cell and dragging down.

A2									
	A	B	C	D	E	F	G	H	I
1	Participant ID	Timestamp	Date:	Quintron Bre	Hydrogen:	Methane:	Carbon Dioxi	Correction:	Comments:
2	789	#####	1/27/19	Quintron 2	18	3	3.7	1.48	
3	789	#####	2/1/19	Quintron 2	26	2	4.8	1.14	
4	789	#####	2/3/19	Quintron 2	1	1	4.4	1.25	
5	789	#####	2/3/19	Quintron 2	1	1	4.6	1.19	
6	789	#####	2/13/19	Quintron 2	72	2	4.5	1.22	
7	789	#####	2/14/19	Quintron 2	63	2	5.1	1.07	
8	789	#####	2/17/19	Quintron 2	25	2	4.6	1.19	
9	U700	####	1/26/19	Quintron 2	0	1	5	1.1	
10		####	1/29/19	Quintron 1	7	1	3.7	1.48	
11		#####	2/12/19	Quintron 2	13	1	4.6	1.19	

Tedious to do? Yup. So...

- Save the corrected dirty_data.csv as CSV format. Close it and open dirty_data_square.csv.

Possible Data Loss Some features might be lost if you save this workbook in the comma-delimited (.csv) format									
A1									
	A	B	C	D	E	F	G	H	I
1	Timestamp	Participant ID	Date:	Quintron Bre	Hydrogen:	Methane:	Carbon Dioxi	Correction:	Comments:
2	#####	U700	1/26/19	Quintron 2	0	1	5	1.1	
3	#####	U739	12/30/99	Quintron 2	0	1	4.3	1.27	
4	#####	U744	1/26/19	Quintron 1	1	1	5.2	1.05	
5	#####	U725	1/26/19	Quintron 1	11	2	3.6	1.52	
6	#####	U788	1/26/19	Quintron 2	0	1	3.8	1.44	
7	#####	U709	1/26/19	Quintron 2	5	1	5.2	1.05	
8	#####	U721	1/26/19	Quintron 1	6	2	2.6	1.52	

We've done the tedious work to make the data (almost) square for you.

5. Standardize the column names. Look at the column names in row 1.

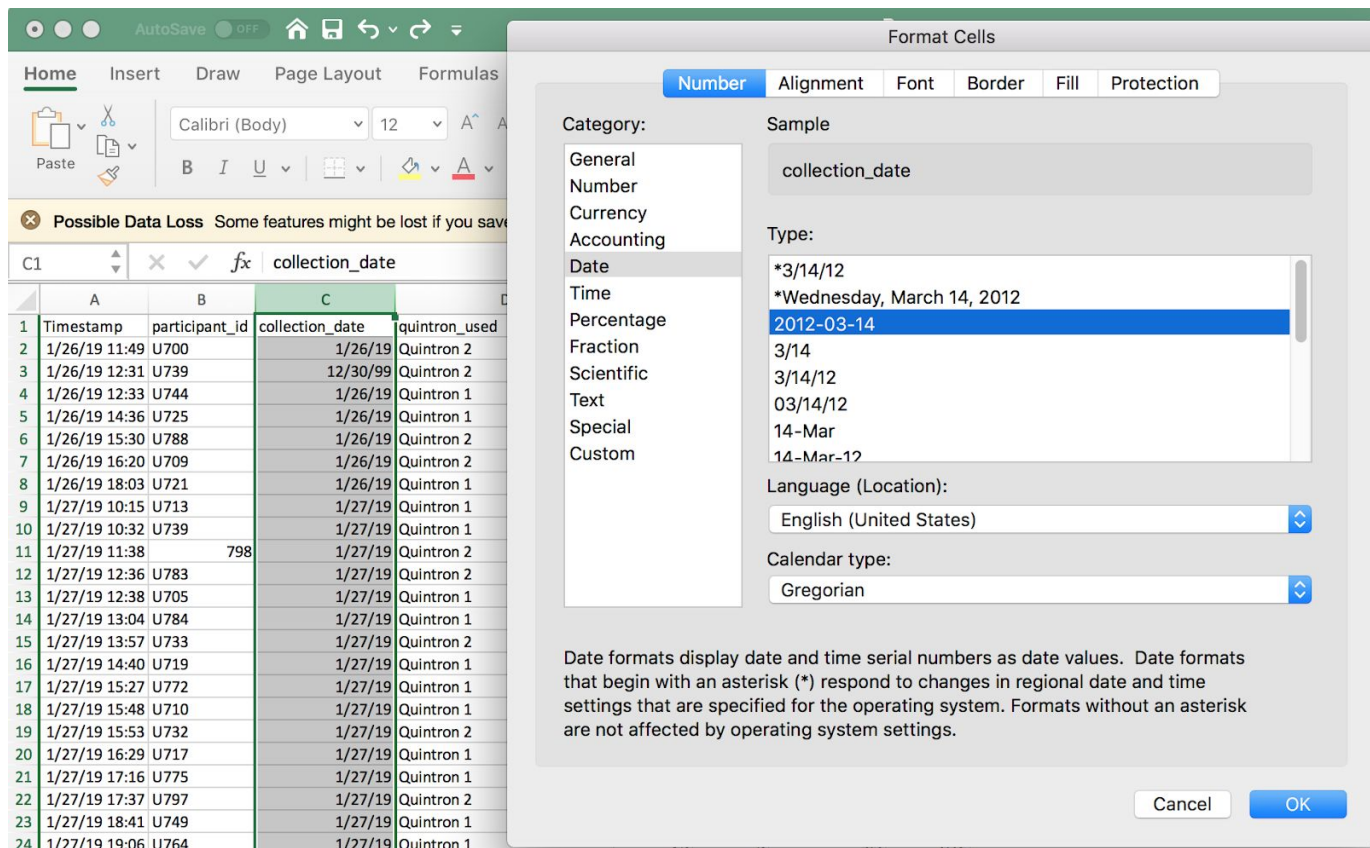
Good column names:

- Have no spaces.
- Have no non-alphanumeric (i.e. special) characters beyond an underscore or hyphens.
- Do *not* mix hyphens and underscores. Pick one (underscores) and use it for the entire spreadsheet.
- Include the units at the end.
- Have no spaces in the name, before or after the name.
- Are short and meaningful without using non-standard abbreviations.

Current name	Better name	Avoid
Timestamp	timestamp	Timestamp:
Participant ID:	participant_id	Part. ID
Date:	collection_date	Coll. date
Quintron BreathTracker Used:	quintron_used	
Hydrogen:	hydrogen_ppm	
Methane:	methane_ppm	
Carbon Dioxide:	carbon_dioxide_percent	CO2 %
Correction:	correction_percent	Corr. %
Comments:	comments	

6. Standardize the dates. Select the 'collection_date' column, and select Format > Cells from the menu. Choose date, and the ISO 8601 standard format.

C1	A	B	C
1	Timestamp	participant_id	collection_date
2	1/26/19 11:49	U700	1/26/19
3	1/26/19 12:31	U739	12/30/99
4	1/26/19 12:33	U744	1/26/19



Click “OK” and allow the standard format to be applied.

All dates in raw data should be in the ISO 8601 standard: YYYY-MM-DD (i.e. 2019-08-30). Do not trust Excel to auto-format dates.

PUBLIC SERVICE ANNOUNCEMENT:

OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.

THIS IS **THE** CORRECT WAY TO WRITE NUMERIC DATES:

2013-02-27

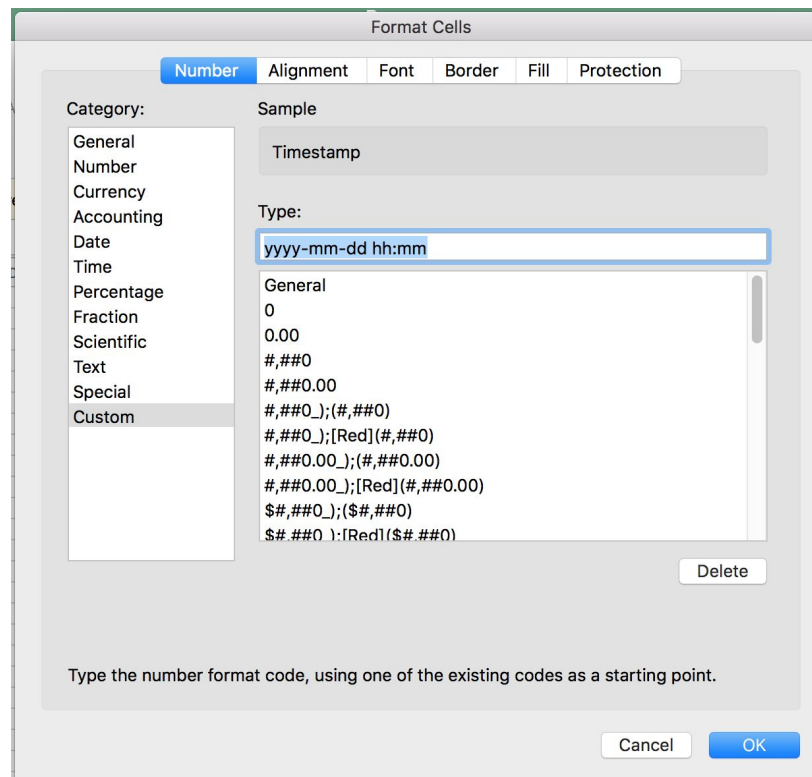
THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:

02/27/2013 02/27/13 27/02/2013 27/02/13
 20130227 2013.02.27 27.02.13 27-02-13
 27.2.13 2013. II. 27. 27²-13 2013.158904109
 MMXIII-II-XXVII MMXIII ^{LVII} 1330300800
 ((3+3)×(111+1)-1)×3/3-1/3³ 2013 Missy
 10/11011/1101 02/27/2013 01237 2013

For timestamps, we should follow a similarly standard format: YYYY-MM-DD HH:MM. Sadly Excel does not even offer this format. Select the “Timestamp” column. Choose Format > Cells from the menu. Choose “Custom”. Put “yyyy-mm-dd hh:mm” into the Type: window. Click OK.

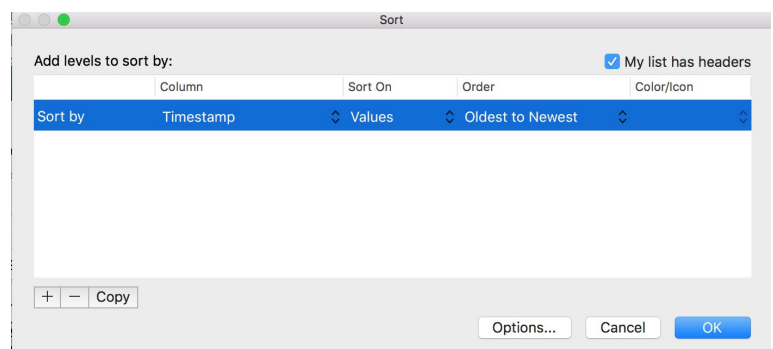
What time zone are these times in? Is daylight savings time included? When designing a study, the safest way to store dates is in UTC or coordinated universal time, with or without a time-zone modifier.

(<https://xkcd.com/1179/>)



	A	B	C	D	E	F	G	H	
1	Timestamp	participant_id	collection_date	quintron_used	hydrogen_	methane_	carbon_dioxide_	correction_	comment
2	2019-01-26 11:49	U700	2019-01-26	Quintron 2	0	1	5	1.1	
3	2019-01-26 12:31	U739	1999-12-30	Quintron 2	0	1	4.3	1.27	
4	2019-01-26 12:33	U744	2019-01-26	Quintron 1	1	1	5.2	1.05	
5	2019-01-26 14:36	U725	2019-01-26	Quintron 1	11	2	3.6	1.52	
6	2019-01-26 15:30	U788	2019-01-26	Quintron 2	0	1	3.8	1.44	
7	2019-01-26 16:20	U709	2019-01-26	Quintron 2	5	1	5.2	1.05	
8	2019-01-26 18:03	U721	2019-01-26	Quintron 1	6	2	3.6	1.52	
9	2019-01-27 10:15	U713	2019-01-27	Quintron 1	1	1	4.6	1.19	

6. Validate values and identify non-standard entries. We can use sorting to see what our values look like for each column. Starting from the left "Timestamp" column, choose Data > Sort



After sorting by that column, look at the values, particularly the first and last values. Do they all make sense? Are the values formatted the same? Are any clearly off?

When sorting by the 'participant_id' note that there are a few entries at the start that are missing their 'U':

B3 fx 798

	A	B	C	
1	Timestamp	participant_id	collection_date	quintro
2	2019-01-27 11:38	798	2019-01-27	Quintro
3	2019-02-01 11:11	798	2019-02-01	Quintro
4	2019-02-03 14:27	798	2019-02-03	Quintro
5	2019-02-03 14:30	798	2019-02-03	Quintro
6	2019-02-13 18:08	798	2019-02-13	Quintro
7	2019-02-14 11:08	798	2019-02-14	Quintro
8	2019-02-17 12:27	798	2019-02-17	Quintro
9	2019-01-26 11:49	U700	2019-01-26	Quintro

Go ahead and add the U so these rows match all of the others.

	A	B	C
1	Timestamp	participant_id	collection_date
2	2019-01-27 11:38	U798	2019-01-27
3	2019-02-01 11:11	U798	2019-02-01
4	2019-02-03 14:27	U798	2019-02-03
5	2019-02-03 14:30	U798	2019-02-03
6	2019-02-13 18:08	U798	2019-02-13
7	2019-02-14 11:08	U798	2019-02-14
8	2019-02-17 12:27	U798	2019-02-17

It is important that all cells in a column are formatted in the same way. Empty or missing cells need also to be carefully considered, and their entire row removed if appropriate.

When sorting by 'hydrogen_ppm', 'methane_ppm', or 'carbon_dioxide_percent', notice the rows where the units are included:

390	2019-02-13 18:08	U798	2019-02-13	Quintron 2	72	2	4.5	1.22
391	2019-02-12 13:58	U751	2019-02-12	Quintron 1	73	4	4.5	1.22
392	2019-02-04 10:00	U730	2019-02-04	Quintron 2	96	40	6.1	0.91
393	2019-01-27 17:16	U775	2019-01-27	Quintron 1	13ppm	2ppm	4.70%	1.17
394	2019-02-14 12:10	U775	2019-02-14	Quintron 1	20ppm	2ppm	3.50%	1.57
395	2019-01-28 17:41	U775	2019-01-28	Quintron 1	5ppm	1ppm	4.30%	1.27

Remove the units, leaving behind only the values.

387	2019-02-03 13:47	U763	2019-02-03	Quintron 1	48	4	4.7	1.17
388	2019-02-19 16:27	U743	2019-02-19	Quintron 2	52	3	3.5	1.57
389	2019-02-14 11:08	U798	2019-02-14	Quintron 2	63	2	5.1	1.07
390	2019-02-13 18:08	U798	2019-02-13	Quintron 2	72	2	4.5	1.22
391	2019-02-12 13:58	U751	2019-02-12	Quintron 1	73	4	4.5	1.22
392	2019-02-04 10:00	U730	2019-02-04	Quintron 2	96	40	6.1	0.91
393	2019-01-27 17:16	U775	2019-01-27	Quintron 1	13	2	4.70%	1.17
394	2019-02-14 12:10	U775	2019-02-14	Quintron 1	20	2	3.50%	1.57
395	2019-01-28 17:41	U775	2019-01-28	Quintron 1	5	1	4.30%	1.27
396								

Units should be the same for an entire column. It is a good practice to include them in the column name.

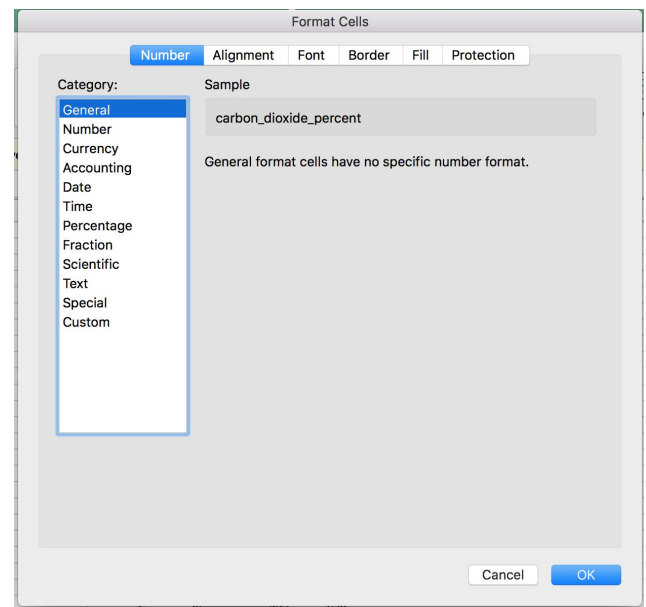
7. Ensure all columns are in the same format. What happened when you tried remove “%” from ‘carbon_dioxide_percent’ cells that had a ‘%’ in them?

390	2019-02-13 18:08	U798	2019-02-13	Quintron 2	72	2	4.5	1.22
391	2019-02-12 13:58	U751	2019-02-12	Quintron 1	73	4	4.5	1.22
392	2019-02-04 10:00	U730	2019-02-04	Quintron 2	96	40	6.1	0.91
393	2019-01-27 17:16	U775	2019-01-27	Quintron 1	13	2	470.00%	1.17
394	2019-02-14 12:10	U775	2019-02-14	Quintron 1	20	2	350.00%	1.57
395	2019-01-28 17:41	U775	2019-01-28	Quintron 1	5	1	430.00%	1.27

That isn't good.

- Be sure the percentage columns are all the same type.
- Select column G (the ‘carbon_dioxide_percent’ column).
- Choose Format > Cell from the menu. Choose the ‘general’ type:

Now all is fine.



387	2019-02-03 13:47	U763	2019-02-03	Quintron 1	48	4	4.7	1.17
388	2019-02-19 16:27	U743	2019-02-19	Quintron 2	52	3	3.5	1.57
389	2019-02-14 11:08	U798	2019-02-14	Quintron 2	63	2	5.1	1.07
390	2019-02-13 18:08	U798	2019-02-13	Quintron 2	72	2	4.5	1.22
391	2019-02-12 13:58	U751	2019-02-12	Quintron 1	73	4	4.5	1.22
392	2019-02-04 10:00	U730	2019-02-04	Quintron 2	96	40	6.1	0.91
393	2019-01-27 17:16	U775	2019-01-27	Quintron 1	13	2	4.7	1.17
394	2019-02-14 12:10	U775	2019-02-14	Quintron 1	20	2	3.5	1.57
395	2019-01-28 17:41	U775	2019-01-28	Quintron 1	5	1	4.3	1.27
396								

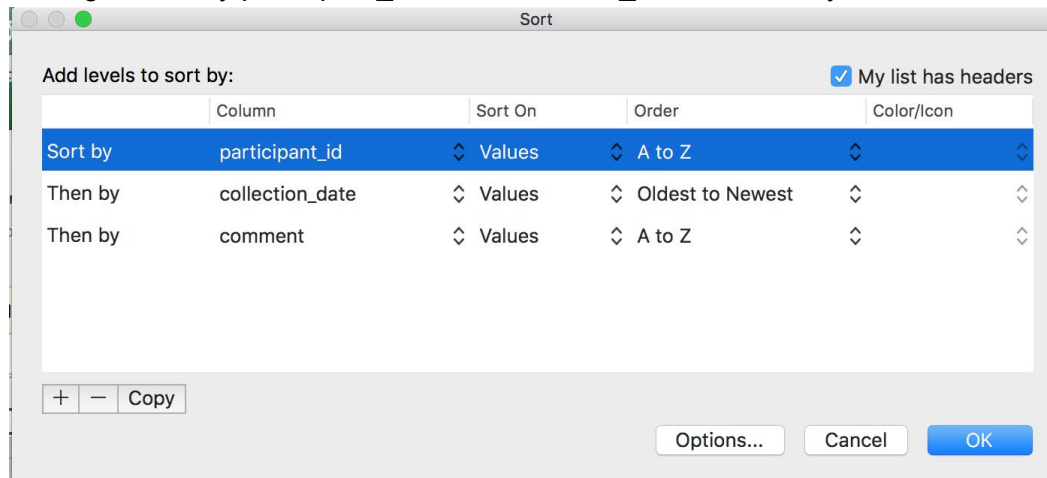
8. Deal with comments. Now look at the comments we have by sorting by the comment column:

comment
2 I took two measurements today because my first measurement's numbers came back much lower from the first week so I wanted t
2 none
2 none
2 Previous breath sample for U709 was an error; forgot to close the valve on the quintron. H2 should be 1 not 0.
7 second try
2 taken at approx 1:04 pm

We can quickly delete the 'none' comments.

We can also delete the 'taken at approx 1:04pm' comment. (collection date is enough).

Sort again first by participant_id, then collection_date, and finally comment



Scroll through to find comments.

39	2019-01-26 16:20 U709	2019-01-26 Quintron 2	5	1	5.2	1.05	
40	2019-01-29 14:36 U709	2019-01-29 Quintron 2	1	1	4.5	1.22	Previous breath sample for U709 was an error; forgot to close the valve on the quintron. H2 should be 1 not 0.
41	2019-01-29 14:30 U709	2019-01-29 Quintron 2	0	1	5.3	1.03	
42	2019-02-01 11:57 U709	2019-02-01 Quintron 2	3	2	5.3	1.03	
43	2019-02-13 11:34 U709	2019-02-13 Quintron 2	3	1	4.9	1.12	
44	2019-02-15 16:01 U709	2019-02-15 Quintron 2	2	1	5.5	1	

For U709 we note the collection from 2019-01-29 has two entries. The comment indicates the other had an error. We can delete the *other* row for this participant for 2019-01-29.

U729 noted the collection on 2019-02-12 had an unexpected value. *What should we do with this comment?*

U781 on 2019-02-13, "I took two measurements today because my first measurement's numbers came back much lower from the first week so I wanted to make sure I was breathing correctly. This second try was fairly similar but still very different than the first week's measurements."

We can delete the second measurement (the one with the comment) for this participant on this day.

What else could we do with this measurement?

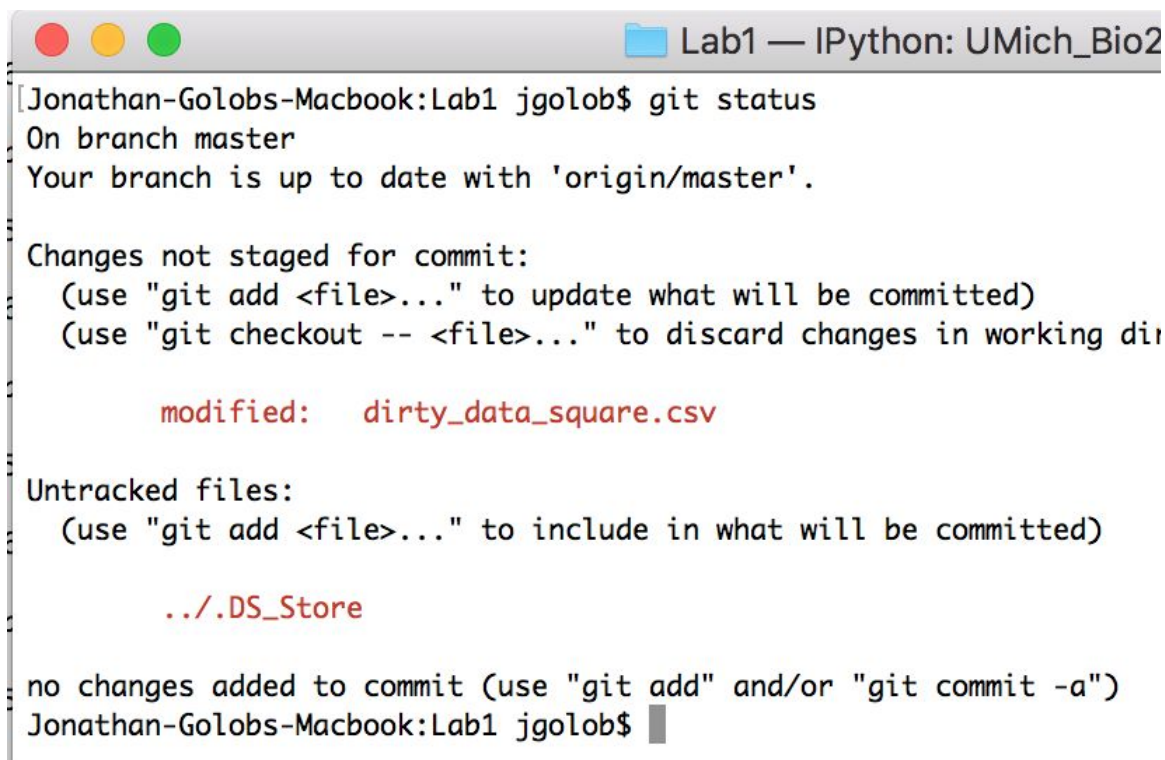
U785 on 2019-02-21 notes 'second try'. *What do you want to do with this comment and row?*

9. Save the file as “dirty_data_square.csv” in CSV format into the “Lab1” folder, overwriting the previous version of this file. Ignore if Excel complains.
10. Use git to see what we changed from the original version.

How can we keep both the original version of the data file and this corrected version? Didn't we just save over the original version with the changes? We can use `git` to help us out here.

Go back to the terminal window. Type in the Lab1 directory:

```
cd ~/Documents/UMich_Bio201_F19/Lab1
git status
```



```

[Jonathan-Golobs-Macbook:Lab1 jgolob$ git status
On branch master
Your branch is up to date with 'origin/master'.

Changes not staged for commit:
  (use "git add <file>..." to update what will be committed)
  (use "git checkout -- <file>..." to discard changes in working dir)

        modified:   dirty_data_square.csv

Untracked files:
  (use "git add <file>..." to include in what will be committed)

        ../.DS_Store

no changes added to commit (use "git add" and/or "git commit -a")
Jonathan-Golobs-Macbook:Lab1 jgolob$
```

Git has noted you have modified the `dirty_data_square.csv` file.

If you want to see what changed you can type,

```
git diff dirty_data_square.csv
```

```

Lab1 — IPython: UMich_Bio201_F19/Lab1 — less • git diff dirty_data_square.csv — 134x30
diff --git a/Lab1/dirty_data_square.csv b/Lab1/dirty_data_square.csv
index a1b6db5..0f01144 100644
--- a/Lab1/dirty_data_square.csv
+++ b/Lab1/dirty_data_square.csv
@@ -1,395 +1,393 @@
-Timestamp,Participant ID:,Date:,Quintron BreathTracker Used:,Hydrogen:,Methane:,Carbon Dioxide:,Correction:,Comments:
-1/26/19 11:49,U700,1/26/19,Quintron 2,0,1,5,1.1,
-1/26/19 12:31,U739,12/30/99,Quintron 2,0,1,4,3,1.27,
-1/26/19 12:33,U744,1/26/19,Quintron 1,1,1,5,2,1.05,
-1/26/19 14:36,U725,1/26/19,Quintron 1,11,2,3,6,1.52,
-1/26/19 15:30,U788,1/26/19,Quintron 2,0,1,3,8,1.44,
-1/26/19 16:20,U709,1/26/19,Quintron 2,5,1,5,2,1.05,
-1/26/19 18:03,U721,1/26/19,Quintron 1,6,2,3,6,1.52,
-1/27/19 10:15,U713,1/27/19,Quintron 1,1,1,4,6,1.19,
-1/27/19 10:32,U739,1/27/19,Quintron 1,6,1,3,8,1.44,
-1/27/19 11:38,798,1/27/19,Quintron 2,18,3,3,7,1.48,
-1/27/19 12:36,U783,1/27/19,Quintron 2,7,2,5,4,1.01,
-1/27/19 12:38,U705,1/27/19,Quintron 1,3,1,4,1,1.34,
-1/27/19 13:04,U784,1/27/19,Quintron 1,3,1,5,1,1.1,
-1/27/19 13:57,U733,1/27/19,Quintron 2,5,1,5,3,1.03,
-1/27/19 14:40,U719,1/27/19,Quintron 1,3,1,4,1,37,
-1/27/19 15:27,U772,1/27/19,Quintron 1,10,2,4,5,1.22,
-1/27/19 15:48,U710,1/27/19,Quintron 1,4,1,4,3,1.27,
-1/27/19 15:53,U732,1/27/19,Quintron 2,6,2,5,5,1.0,
-1/27/19 16:29,U717,1/27/19,Quintron 1,23,2,4,5,1.22,none
-1/27/19 17:16,U775,1/27/19,Quintron 1,13ppm,2ppm,4.70%,1.17,
-1/27/19 17:37,U797,1/27/19,Quintron 2,11,2,5,1,1.07,
-1/27/19 18:41,U749,1/27/19,Quintron 1,3,3,3,6,1.52,
-1/27/19 19:06,U764,1/27/19,Quintron 1,25,3,4,9,1.12,

```

Git has kept track of the original version and all of our modifications. Type `q` to quit.

11. Commit our modified document with git.

In order to keep our changes to `dirty_data_square.csv`, we need to tell git to commit those changes. We add files we want to commit, then `commit` them with a message.

In the Lab1 directory in the terminal, type:

```

git add dirty_data_square.csv
git commit -m corrected

```

```

Lab1 — IPython: UMich_Bio201_F19/Lab1
[Jonathan-Golobs-Macbook:Lab1 jglob$ git add dirty_data_square.csv
[Jonathan-Golobs-Macbook:Lab1 jglob$ git commit -m corrected
[master 154dfd3] corrected
 1 file changed, 393 insertions(+), 395 deletions(-)
 rewrite Lab1/dirty_data_square.csv (99%)
Jonathan-Golobs-Macbook:Lab1 jglob$

```

Every commit needs a message, something to say what changed. Here we used a shortcut, `-m`. Everything after `-m` is used as the message.

Now git has been told to keep our changes in the `dirty_data_square.csv` file.

Let's look at the log to get a sense of what is going on here. In the terminal type:

```
git log
```

```

Lab1 — IPython: UMich_Bio201_F19/Lab1 — less ◀ git log

commit 154dfd3ca1ae1beeaba8ea42a540310bd0a6af93 (HEAD -> master)
Author: Jonathan Golob <j-dev@golob.org>
Date:   Tue Sep 3 10:27:36 2019 -0400

    corrected

commit 0e844035ddfbf4a7857d916a21bec412b274c525 (tag: v1.0, origin/master, origin/HEAD)
Author: Jonathan Golob <j-dev@golob.org>
Date:   Tue Sep 3 10:18:32 2019 -0400

    Pre correction data
  
```

We can see a few commits here. A commit tagged as `v1.0` that was the pre-correction data and our new commit with the corrected data. Our message was saved. To exit the log type, `q`

We can add a tag to make it easier to find this version of the data. In the terminal, type:

```
git tag corrected
```

```
git log
```

```

Lab1 — IPython: UMich_Bio201_F19/Lab1 — less ◀ git log -

commit 154dfd3ca1ae1beeaba8ea42a540310bd0a6af93 (HEAD -> master, tag: corrected)
Author: Jonathan Golob <j-dev@golob.org>
Date:   Tue Sep 3 10:27:36 2019 -0400

    corrected

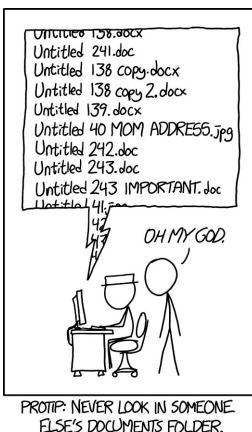
commit 0e844035ddfbf4a7857d916a21bec412b274c525 (tag: v1.0, origin/master, origin/HEAD)
Author: Jonathan Golob <j-dev@golob.org>
Date:   Tue Sep 3 10:18:32 2019 -0400

    Pre correction data
  
```

We can now see our tag, `corrected`.

Why are we doing this? The alternative is to change the file name each time we modify the file, which is messy and prone to error (i.e. accidentally overwriting a version we wanted to keep). git naturally keeps track of all the changes in a file, and even lets us go back in time.

<https://xkcd.com/1459/>

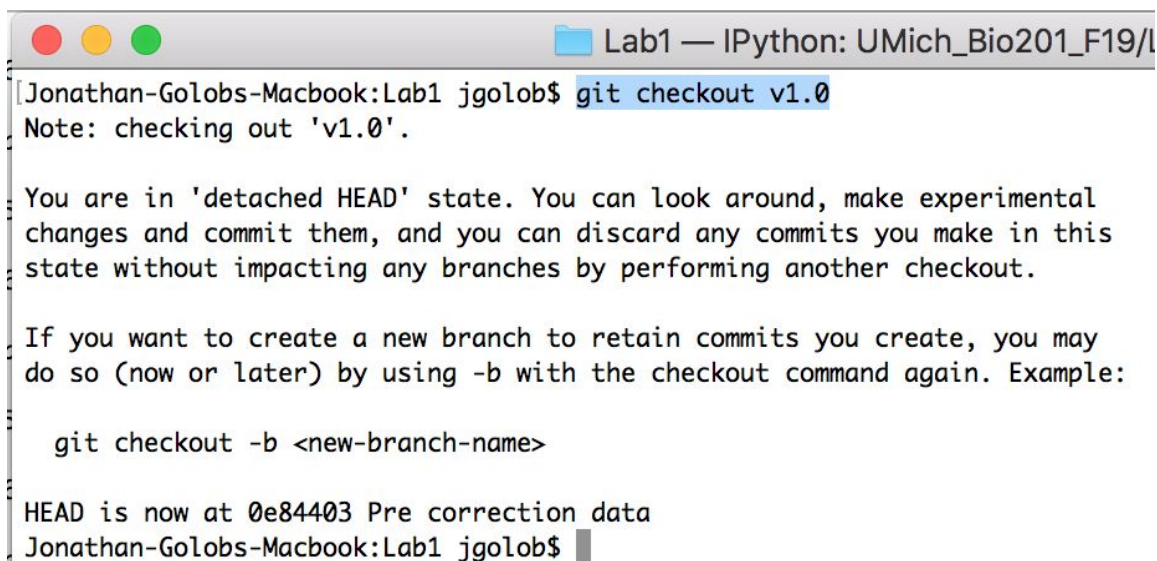


12. Use git to jump between the new and old versions

Be sure `dirty_data_square.csv` is closed in Excel.

In the terminal window in the Lab1 directory, type:

```
git checkout v1.3
```



```
Jonathan-Golobs-Macbook:Lab1 jgolob$ git checkout v1.0
Note: checking out 'v1.0'.

You are in 'detached HEAD' state. You can look around, make experimental
changes and commit them, and you can discard any commits you make in this
state without impacting any branches by performing another checkout.

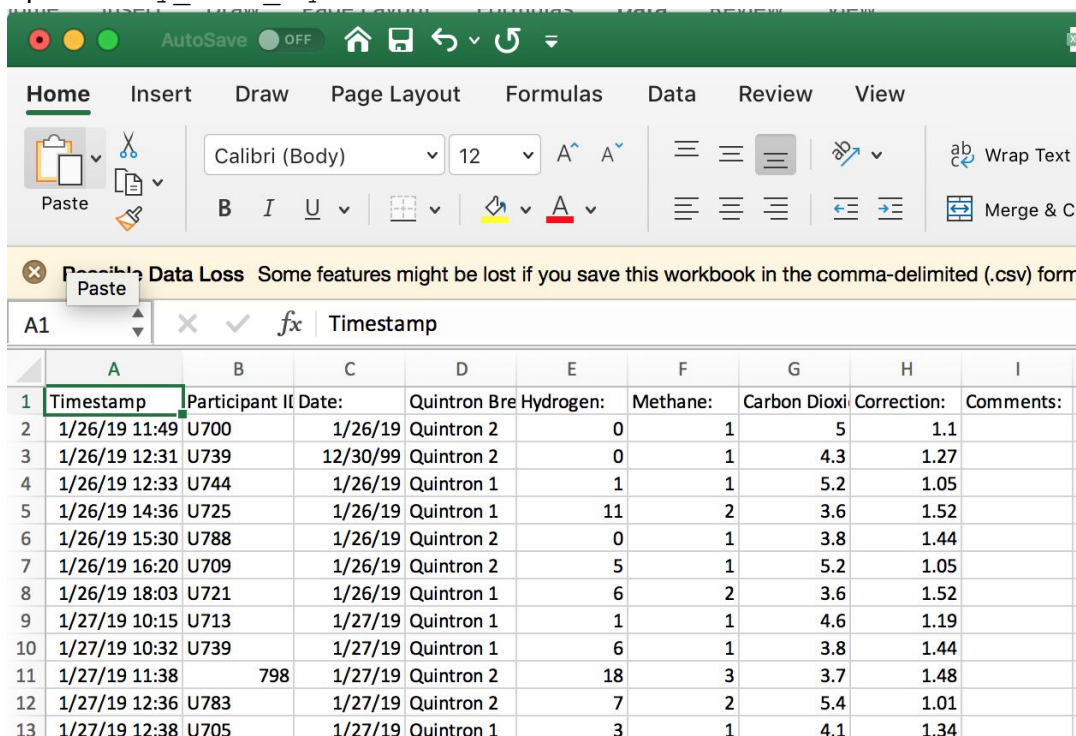
If you want to create a new branch to retain commits you create, you may
do so (now or later) by using -b with the checkout command again. Example:

  git checkout -b <new-branch-name>

HEAD is now at 0e84403 Pre correction data
Jonathan-Golobs-Macbook:Lab1 jgolob$
```

The *'detached head'* warning is git is warning you that we have gone back in time.

Open `dirty_data_square.csv` in Excel.



Warning: Possible Data Loss. Some features might be lost if you save this workbook in the comma-delimited (.csv) format.

	A	B	C	D	E	F	G	H	I
1	Timestamp	Participant ID	Date:	Quintron Bre	Hydrogen:	Methane:	Carbon Dioxi	Correction:	Comments:
2	1/26/19 11:49	U700	1/26/19	Quintron 2	0	1	5	1.1	
3	1/26/19 12:31	U739	12/30/99	Quintron 2	0	1	4.3	1.27	
4	1/26/19 12:33	U744	1/26/19	Quintron 1	1	1	5.2	1.05	
5	1/26/19 14:36	U725	1/26/19	Quintron 1	11	2	3.6	1.52	
6	1/26/19 15:30	U788	1/26/19	Quintron 2	0	1	3.8	1.44	
7	1/26/19 16:20	U709	1/26/19	Quintron 2	5	1	5.2	1.05	
8	1/26/19 18:03	U721	1/26/19	Quintron 1	6	2	3.6	1.52	
9	1/27/19 10:15	U713	1/27/19	Quintron 1	1	1	4.6	1.19	
10	1/27/19 10:32	U739	1/27/19	Quintron 1	6	1	3.8	1.44	
11	1/27/19 11:38	798	1/27/19	Quintron 2	18	3	3.7	1.48	
12	1/27/19 12:36	U783	1/27/19	Quintron 2	7	2	5.4	1.01	
13	1/27/19 12:38	U705	1/27/19	Quintron 1	3	1	4.1	1.34	

It's the original uncorrected version!

Close `dirty_data_square.csv` in Excel. Go back to the terminal and type:
`git checkout master`

Open `dirty_data_square.csv` in Excel.

⌘ Possible Data Loss Some features might be lost if you save this workbook in the comma-delimited (.csv) format

	A	B	C	D	E	F	G	H	I
1	Timestamp	participant_i	collection_d	quintron_use	hydrogen_pp	methane_pp	carbon_dioxi	correction_p	comment
2	1/26/19 11:49	U700	1/26/19	Quintron 2	0	1	5	1.1	
3	1/29/19 12:11	U700	1/29/19	Quintron 1	7	1	3.7	1.48	
4	2/12/19 17:10	U700	2/12/19	Quintron 2	13	1	4.6	1.19	
5	2/14/19 17:10	U700	2/14/19	Quintron 1	3	0	3.9	1.41	
6	1/29/19 13:23	U701	1/29/19	Quintron 2	10	12	4.5	1.22	
7	2/5/19 21:29	U701	2/5/19	Quintron 2	45	8	5.3	1.03	
8	2/16/19 15:36	U701	2/16/19	Quintron 1	5	5	5.4	1.01	
9	2/21/19 12:18	U701	2/21/19	Quintron 1	6	6	3.9	1.41	
10	2/4/19 12:00	U702	2/3/19	Quintron 1	3	0	3.7	1.48	
11	2/4/19 12:02	U702	2/4/19	Quintron 1	3	0	3.7	1.48	

We are back to our corrected version.

Note: Excel ruined the format of the dates again. For now we will not repeat the steps we did above. This is another reason it is better to work with data with tools (like RStudio) other than Excel.