Using spreadsheet programs for scientific data

Spreadsheet programs are use very useful graphical interfaces for designing data tables and handling very basic data quality control functions.

The cardinal rules of using spreadsheet programs for data:

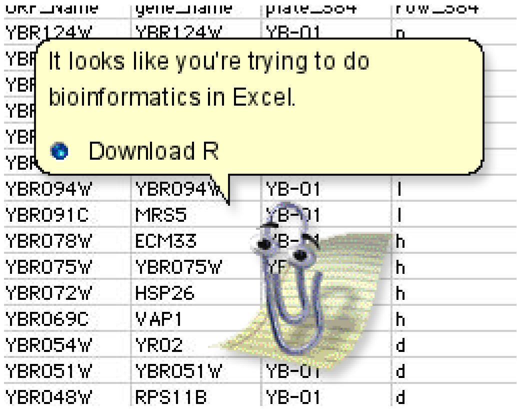
1. Put all your fields in columns
2. Put each observation in its own row
3. Don't @#$%! with it
4. Export to CSV.

In reality, though, many scientists use spreadsheet programs for much more than this. We use them to create data tables for publications, to generate summary statistics, and make figures. Generating tables for publications in a spreadsheet is not optimal- often, when formatting a data table for publication, we're reporting key summary statistics in a way that is not really meant to be read as data, and often involves special formatting (merging cells, creating borders, making it pretty). We advise you to do this sort of operation within your document editing software.

The latter two applications, generating statistics and figures, should be used with caution: because of the graphical, drag and drop nature of spreadsheet programs, it can be very difficult, if not impossible, to replicate your steps, particularly if your stats or figures require you to do more complex calculations. Furthermore, in doing calculations in a spreadsheet, it's easy to accidentally apply a slightly different formula to multiple adjacent cells. When using a command-line based statistics program like R or SAS, it's practically impossible to accidentally apply a calculation to one observation in your dataset but not another unless you're doing it on purpose.

HOWEVER, there are circumstances where you might want to use a spreadsheet program to produce "quick and dirty" calculations or figures, and some of these features can be used in data cleaning, prior to importation into a statistical analysis program. We will show you how to use some features of spreadsheet programs to check your data quality along the way and produce preliminary summary statistics.

In this lesson, we will assume that you are most likely using Excel as your primary spreadsheet program- there are others (gnumeric, Calc from OpenOffice), and their functionality is similar, but Excel seems to be the program most used by biologists and ecologists.



ORF_Name	gene_name	plate_384	row_384
YBR124W	YBR124W	YB-01	n
YBR			
YBR			
YBR			
YBR			
YBR			
YBR094W	YBR094W	YB-01	l
YBR091C	MRS5	YB-01	l
YBR078W	ECM33	YB-0	h
YBR075W	YBR075W	YB	h
YBR072W	HSP26		h
YBR069C	VAP1		h
YBR054W	YRO2		d
YBR051W	YBR051W	YB-01	d
YBR048W	RPS11B	YB-01	d

It looks like you're trying to do bioinformatics in Excel.

● Download R

by @tim_yates

In this lesson, we're going to talk about:

1. Formatting data tables in spreadsheets
2. Common formatting mistakes by spreadsheet users
    a. Dates as data
3. Basic quality control and data manipulation in spreadsheets
4. Exporting data from spreadsheets

Formatting data tables in Spreadsheets

The most common mistake a casual spreadsheet user makes is by treating the program like it is a lab notebook- that is, relying on context, notes in the margin, spatial layout of data and fields to convey information. As humans, we can (usually) interpret these things, but computers are dumb, and unless we explain to the computer what every single thing means, it will not be able to see how our data fit together.

This is why it's extremely important to set up well-formatted tables from the outset- before you even start entering data from your very first preliminary experiment. Spreadsheets are powerful because they allow us to connect things that relate to each other in a machine-readable way. Failing to use this functionality is essentially the same as scanning your datasheets and saving them as jpegs- digitally backed up, but not really doing anything for anybody. When you don't set up your spreadsheet which allows the computer to see how things are connected, you're either creating a lot of work for you or for someone else, or dooming your data to obscurity. There are two simple rules you should keep in mind when entering your data into any spreadsheet:

- Each data cell is an observation that must have all the relevant information connected to it for it to stand on its own.

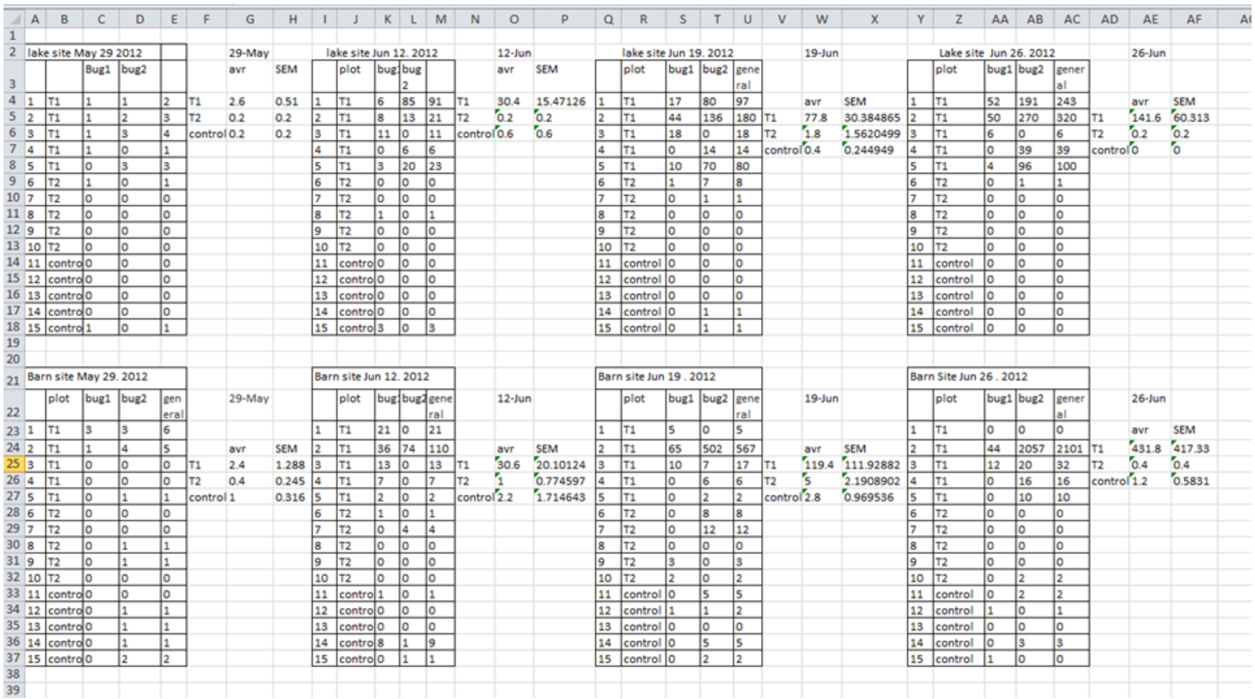
- You must make it clear to the computer how the data cells relate to the relevant information and each other.

So, you can see how these two points a
pply directly to how you set up your spreadsheets.

The rule of thumb, when setting up a datasheet, is columns= fields, rows = observations, cells=data.

[create a basic datasheet in Excel]

A common grad student error I've seen is students that create multiple data tables within one spreadsheet. NEVER DO THIS. When you create multiple tables within one spreadsheet, you're drawing false associations between things for the computer, which sees each row as an observation. You're also potentially using the same field name in multiple places, which will make it harder to clean your data up into a usable form



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1																
2	lake site May 29 2012						29-May		lake site Jun 12. 2012						12-Jun	
3			Bug1	bug2			avr	SEM		plot	bug1	bug 2			avr	SEM
4	1	T1	1	1	2	T1	2.6	0.51	1	T1	6	85	91	T1	30.4	15.47126
5	2	T1	1	2	3	T2	0.2	0.2	2	T1	8	13	21	T2	0.2	0.2
6	3	T1	1	3	4	control	0.2	0.2	3	T1	11	0	11	control	0.6	0.6
7	4	T1	1	0	1				4	T1	0	6	6			
8	5	T1	0	3	3				5	T1	3	20	23			
9	6	T2	1	0	1				6	T2	0	0	0			
10	7	T2	0	0	0				7	T2	0	0	0			
11	8	T2	0	0	0				8	T2	1	0	1			
12	9	T2	0	0	0				9	T2	0	0	0			
13	10	T2	0	0	0				10	T2	0	0	0			
14	11	contro	0	0	0				11	contro	0	0	0			
15	12	contro	0	0	0				12	contro	0	0	0			
16	13	contro	0	0	0				13	contro	0	0	0			
17	14	contro	0	0	0				14	contro	0	0	0			
18	15	contro	1	0	1				15	contro	3	0	3			
19																
20																
21	Barn site May 29. 2012								Barn site Jun 12. 2012							
22		plot	bug1	bug2	gen eral		29-May			plot	bug1	bug2	gene ral		12-Jun	
23	1	T1	3	3	6				1	T1	21	0	21			
24	2	T1	1	4	5		avr	SEM	2	T1	36	74	110		avr	SEM
25	3	T1	0	0	0	T1	2.4	1.288	3	T1	13	0	13	T1	30.6	20.10124
26	4	T1	0	0	0	T2	0.4	0.245	4	T1	7	0	7	T2	1	0.774597
27	5	T1	0	1	1	control	1	0.316	5	T1	2	0	2	control	2.2	1.714643
28	6	T2	0	0	0				6	T2	1	0	1			
29	7	T2	0	0	0				7	T2	0	4	4			
30	8	T2	0	1	1				8	T2	0	0	0			
31	9	T2	0	1	1				9	T2	0	0	0			
32	10	T2	0	0	0				10	T2	0	0	0			
33	11	contro	0	0	0				11	contro	1	0	1			
34	12	contro	0	1	1				12	contro	0	0	0			
35	13	contro	0	1	1				13	contro	0	0	0			
36	14	contro	0	1	1				14	contro	8	1	9			
37	15	contro	0	2	2				15	contro	0	1	1			
38																
39																

	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF
1																
2		lake site Jun 19. 2012					19-Jun			Lake site Jun 26. 2012					26-Jun	
3		plot	bug1	bug2	gene ral					plot	bug1	bug2	gener al			
4	1	T1	17	80	97		avr	SEM	1	T1	52	191	243		avr	SEM
5	2	T1	44	136	180	T1	77.8	30.384865	2	T1	50	270	320	T1	141.6	60.313
6	3	T1	18	0	18	T2	1.8	1.5620499	3	T1	6	0	6	T2	0.2	0.2
7	4	T1	0	14	14	control	0.4	0.244949	4	T1	0	39	39	control	0	0
8	5	T1	10	70	80				5	T1	4	96	100			
9	6	T2	1	7	8				6	T2	0	1	1			
10	7	T2	0	1	1				7	T2	0	0	0			
11	8	T2	0	0	0				8	T2	0	0	0			
12	9	T2	0	0	0				9	T2	0	0	0			
13	10	T2	0	0	0				10	T2	0	0	0			
14	11	control	0	0	0				11	control	0	0	0			
15	12	control	0	0	0				12	control	0	0	0			
16	13	control	0	0	0				13	control	0	0	0			
17	14	control	0	1	1				14	control	0	0	0			
18	15	control	0	1	1				15	control	0	0	0			
19																
20																
21	Barn site Jun 19 . 2012								Barn Site Jun 26 . 2012							
22		plot	bug1	bug2	gene ral		19-Jun			plot	bug1	bug2	gener al		26-Jun	
23	1	T1	5	0	5				1	T1	0	0	0		avr	SEM
24	2	T1	65	502	567		avr	SEM	2	T1	44	2057	2101	T1	431.8	417.33
25	3	T1	10	7	17	T1	119.4	111.92882	3	T1	12	20	32	T2	0.4	0.4
26	4	T1	0	6	6	T2	5	2.1908902	4	T1	0	16	16	control	1.2	0.5831
27	5	T1	0	2	2	control	2.8	0.969536	5	T1	0	10	10			
28	6	T2	0	8	8				6	T2	0	0	0			
29	7	T2	0	12	12				7	T2	0	0	0			
30	8	T2	0	0	0				8	T2	0	0	0			
31	9	T2	3	0	3				9	T2	0	0	0			
32	10	T2	2	0	2				10	T2	0	2	2			
33	11	control	0	5	5				11	control	0	2	2			
34	12	control	1	1	2				12	control	1	0	1			
35	13	control	0	0	0				13	control	0	0	0			
36	14	control	0	5	5				14	control	0	3	3			
37	15	control	0	2	2				15	control	1	0	0			
38																
39																

[create an example of this in Excel]

But what about worksheet tabs? that seems like an easy way to organize data, right? Well, yes and no. When you create extra tabs, you fail to allow the computer to see connections in the data that are there. Say, for instance, you make a separate tab for each day you take a measurement.

[create an example of this in Excel]

This is bad practice for two reasons: 1) you are more likely to accidentally add inconsistencies to your data if each time you take a measurement, you start recording data in a new tab, and 2) even if you manage to prevent all inconsistencies from creeping in, you will add an extra step for yourself before you analyze the data because you will have to combine these data into a single datatable. You will have to explicitly tell the computer how to combine tabs- and if the tabs are inconsistently formatted, you might even have to do it by hand!

The next time you're entering data, and you go to create another tab or table, I want you to ask yourself "Self, could I avoid adding this tab by adding another column to my original spreadsheet?"

Your data sheet might get very long over the course of experiment. This makes it harder to enter data if you can't see your headers at the top of the spreadsheet. But do NOT repeat headers. These can easily get mixed into the data, leading to problems down the road.

[Show an example of how to freeze panes in Excel]

Common mistakes by spreadsheet users

Not filling in zeroes

-example: when entering count data for a community, nonzero observations may be rare- why bother if they're mostly zeroes?

> Spreadsheets and statistical programs will likely mis-interpret blank cells that are meant to be zero. This is equivalent to leaving out data. Zero observations are real data! Leaving zero data blank is not good in a written lab notebook, but NEVER okay when you move your data into a digital format
>
> [create an example of this in Excel]

Using bad null values

-example: using -999 or other numerical values (or zero)

> Many statistical programs will not recognize that numeric values of null are indeed null. It will depend on the final application of your data and how you intend to analyse it, but it is essential to use a clearly defined and CONSISTENT null indicator. Blanks (most applications) and NA (for R) are good choices.
>
> **From White et al, 2013, Nine simple ways to make it easier to (re)use your data. Ideas in Ecology and Evolution:**

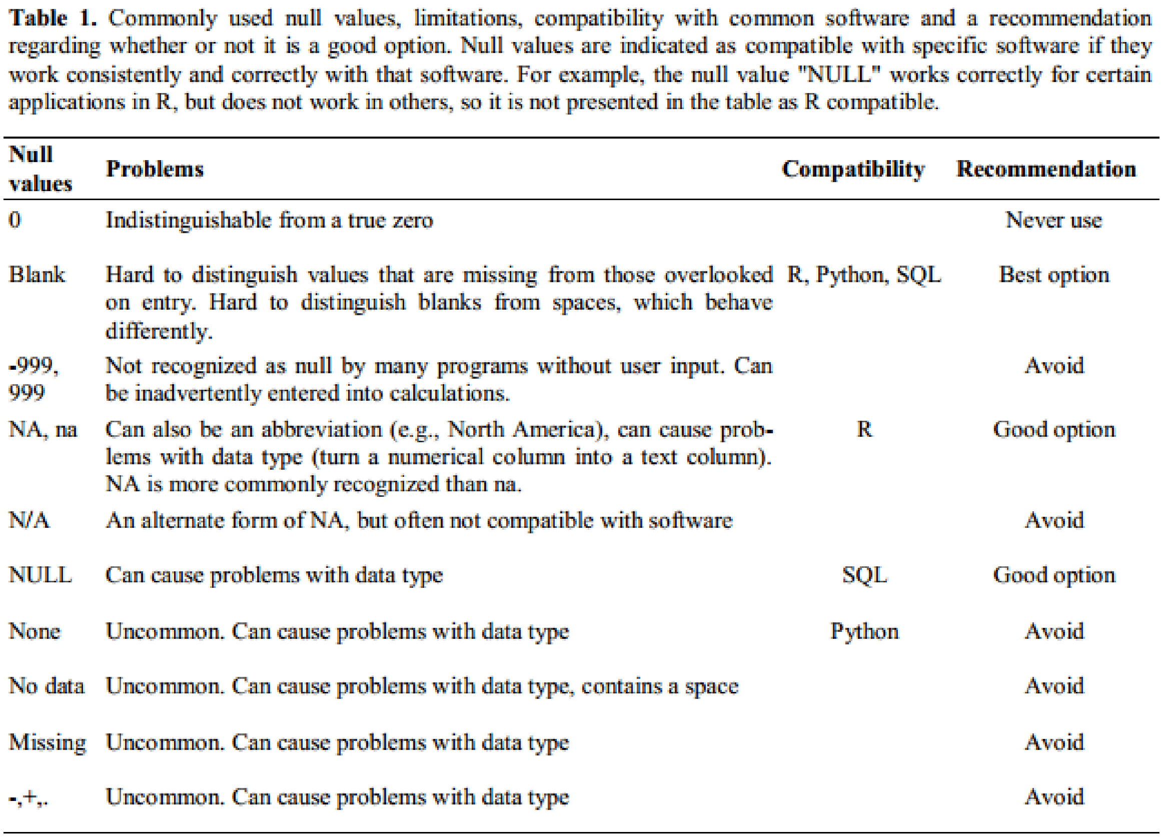


Table 1. Commonly used null values, limitations, compatibility with common software and a recommendation regarding whether or not it is a good option. Null values are indicated as compatible with specific software if they work consistently and correctly with that software. For example, the null value "NULL" works correctly for certain applications in R, but does not work in others, so it is not presented in the table as R compatible.

Null values	Problems	Compatibility	Recommendation
0	Indistinguishable from a true zero		Never use
Blank	Hard to distinguish values that are missing from those overlooked on entry. Hard to distinguish blanks from spaces, which behave differently.	R, Python, SQL	Best option
-999, 999	Not recognized as null by many programs without user input. Can be inadvertently entered into calculations.		Avoid
NA, na	Can also be an abbreviation (e.g., North America), can cause problems with data type (turn a numerical column into a text column). NA is more commonly recognized than na.	R	Good option
N/A	An alternate form of NA, but often not compatible with software		Avoid
NULL	Can cause problems with data type	SQL	Good option
None	Uncommon. Can cause problems with data type	Python	Avoid
No data	Uncommon. Can cause problems with data type, contains a space		Avoid
Missing	Uncommon. Can cause problems with data type		Avoid
-,+,.	Uncommon. Can cause problems with data type		Avoid

Using formatting to convey information

Example: highlighting cells, rows or columns that should be excluded from an analysis, leaving blank rows to indicate separations in data

solution: create a new field to encode which data should be excluded

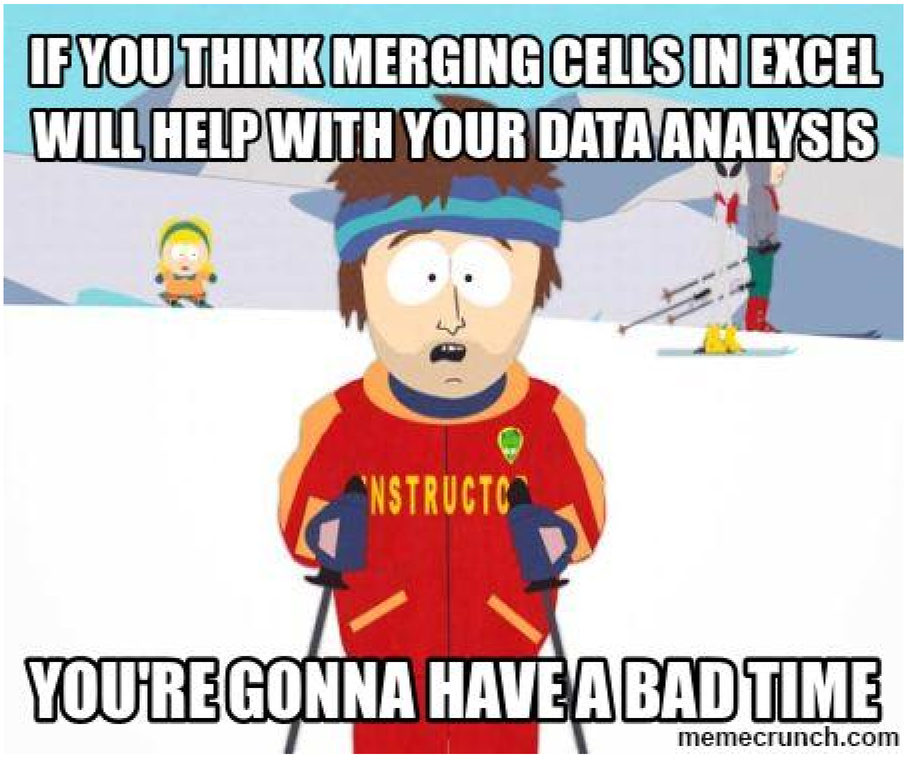
[create an example of this in Excel]

Using formatting to make the data sheet look pretty

Example: merging cells

If you’re not careful, formatting a worksheet to be more aesthetically pleasing can compromise your computer’s ability to see associations in the data. Merged cells are an absolute formatting NO-NO if you want to make your data readable by statistics software. Consider restructuring your data in such a way that you will not need to merge cells to organize your data

[create an example of this in Excel]


IF YOU THINK MERGING CELLS IN EXCEL WILL HELP WITH YOUR DATA ANALYSIS
INSTRUCTOR
YOU'RE GONNA HAVE A BAD TIME
memecrunch.com

by @naupakaz

Placing comments or units in cells

Example: your data was collected, in part, by a summer student you later found out was mis-identifying some of your species, some of the time. You want a way to note these data are suspect.

> Solution: most statistical programs can't see Excel's comments, and would be confused by comments placed within your data cells. Create another field if you need to add notes to cells. Similarly, don't include units- ideally, all the measurements you place in one column should be in the same unit, but if for some reason they aren't, create another field and specify the units the cell is in.
>
> [create an example of this in Excel]

More than one piece of information in a cell

Example: you find one male, and one female of the same species. You enter this as 1M, 1F.

> Solution: Never, ever, EVER include more than one piece of information in a cell. If you need both these measurements, design your data sheet to include this information.
>
> [create an example of this in Excel]

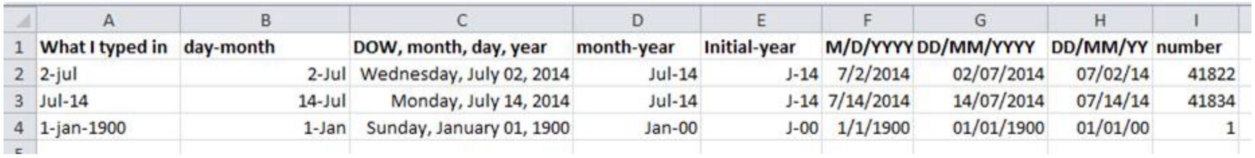
Field name problems

Choose descriptive field names, but be careful not to include: spaces, numbers, or special characters of any kind. Spaces can be misinterpreted and some programs don't like field names that are text strings that start with numbers.

[work through good examples of field names]

Dates as data

Spreadsheet programs have numerous "useful features" which allow them to "handle" dates in a variety of ways.



	A	B	C	D	E	F	G	H	I
1	**What I typed in**	**day-month**	**DOW, month, day, year**	**month-year**	**Initial-year**	**M/D/YYYY**	**DD/MM/YYYY**	**DD/MM/YY**	**number**
2	2-jul	2-Jul	Wednesday, July 02, 2014	Jul-14	J-14	7/2/2014	02/07/2014	07/02/14	41822
3	Jul-14	14-Jul	Monday, July 14, 2014	Jul-14	J-14	7/14/2014	14/07/2014	07/14/14	41834
4	1-jan-1900	1-Jan	Sunday, January 01, 1900	Jan-00	J-00	1/1/1900	01/01/1900	01/01/00	1

But these 'features' often allow ambiguity to creep into your data. Ideally, data should be as unambiguous as possible. The first thing you need to know is that Excel stores dates as a number- see the last column in the above figure. Essentially, it counts the days from a default of December 31, 1899, and thus stores July 2, 2014 as the serial number 41822.

(But wait. That's the default on my version of Excel. We'll get into how this can introduce problems down the line later in this lesson.)

This serial number thing can actually be useful in some circumstances. Say you had a sampling plan where you needed to sample every thirty seven days. In another cell, you could type:

=B2+37

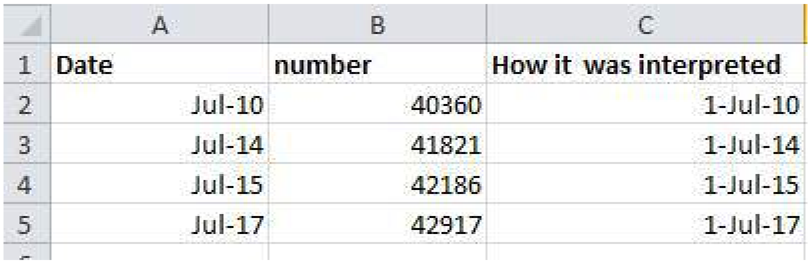
And it would return

8-Aug

because it understands the date as a number 41822, and 41822 +37 =41859 which Excel interprets as August 8, 2014. It retains the format (for the most part) of the cell that is being operated upon, (unless you did some sort of formatting to the cell before, and then all bets are of)

Which brings us to the many 'wonderful' customizations Excel provides in how it displays dates. If you refer to the figure above, you'll see that there's many, MANY ways that ambiguity creeps into your data depending on the format you chose when you enter your data, and if you're not fully cognizant of which format you're using, you can end up actually entering your data in a way that Excel will heavily misinterpret.

Once, I received a dataset from a colleague representing insect counts that were taken every few days over the summer, and things went something like this:



	A	B	C
1	Date	number	How it was interpreted
2	Jul-10	40360	1-Jul-10
3	Jul-14	41821	1-Jul-14
4	Jul-15	42186	1-Jul-15
5	Jul-17	42917	1-Jul-17

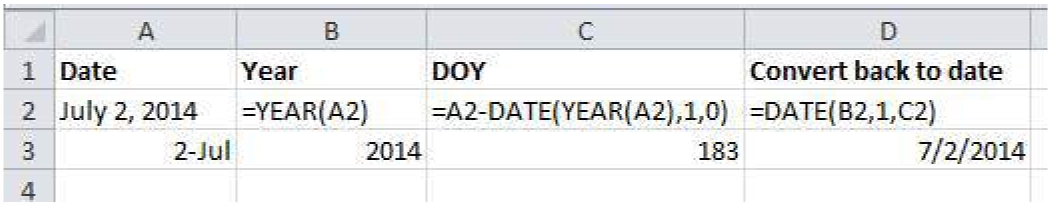
If Excel was to be believed, my colleague had been collecting bugs IN THE FUTURE. Now, I have no doubt this person is highly capable, but I believe time travel was beyond even his grasp.

Thus, in dealing with dates in spreadsheets, we recommend separating date data into separate fields, which will eliminate any chance of ambiguity.

In my own work, I tend to store my dates in two fields: year, and day of year (DOY). Why? Because this is what's useful to me, and there is practically no possibility for ambiguity creeping in.

The types of statistical models I build usually incorporate year as a factor, to account for year-to-year variation, and then I use DOY to measure the passage of time within a year.

So, can you convert all your dates into DOY format? Well, in excel, here's a handy dandy guide:



	A	B	C	D
1	Date	Year	DOY	Convert back to date
2	July 2, 2014	=YEAR(A2)	=A2-DATE(YEAR(A2),1,0)	=DATE(B2,1,C2)
3	2-Jul	2014	183	7/2/2014
4				

[Exercise: pulling month out of Excel dates]

Basic quality control and data manipulation in spreadsheets

When you have a well-structured data table, you can use several simple techniques within your spreadsheet to ensure the data you've entered is free of errors.

*Tip! before doing any quality control operations, ensure your data is stored as “values” and not as formulas. Because formulas refer to other cells, and you may be moving cells around, you may compromise the integrity of your data if you do not take this step!

[Example: converting all data to values: use soybean aphid suction trap dataset for this section]

Sorting

Bad values often sort to bottom or top of column. Sort your data by each field, one at a time. Scan through each column, but pay the most attention to the top and the bottom of a column. If your dataset is well-structured and does not contain formulas, sorting should never affect the integrity of your dataset.

[Example: sorting]

Conditional formatting

Use with caution! But a great way to flag inconsistent values when entering data.

[Example: conditional formatting]

Check on cell formats

A good way to check if you've got data of the wrong type in a column is by checking column format. This can also help prevent issues when you export your data.

[Example: variable format]

Pivot tables

Pivot tables are a very powerful tool in Excel. They're useful to check for issues with data integrity because they provide a quick, visual way to spot things that are amiss, including with categorical variables. They are also great for reshaping data and obtaining summary statistics quickly in a drag and drop interface.

[Example in Pivot Table]

Why would I need to reshape my data? Different analyses require data to be in different formats- example: taking a species list to a diversity analysis.

[Use GLBRC Bees species list example]

Note: these operations can be done in most statistical or programming packages (i.e. using reshape2, plyr in R)

Exporting data from spreadsheets

Storing data in Excel is a bad idea. Why? Because it is a proprietary format, and it is possible that in the future, technology won't exist (or will become sufficiently rare) to make it inconvenient, if not impossible, to open the file.

Think about zipdisks. How many old theses in your lab are "backed up" and stored on zipdisks? Ever wanted to pull out the raw data from one of those?
Exactly.

But more insidious- different versions of Excel may be changed so they handle data differently, leading to inconsistencies.

As an example, do you remember how we talked about how Excel stores dates earlier? Turns out there are multiple defaults for different versions of the software. And you can switch between them all willy-nilly. So… say you're compiling Excel-stored data from multiple sources. There's dates in each file- Excel interprets them as their own internally consistent serial numbers. When you combine the data, Excel will take the serial number from the place you're importing it from, and interpret it using the rule set for the version of Excel you're using. Essentially, you could be adding a huge error to your data, and it wouldn't necessarily be flagged by any data cleaning methods if your ranges overlap.

Storing data in a universal, open, static format will help deal with this problem. Try tab-delimited or csv (more common). Most spreadsheet programs can save to delimited text formats like csv easily, although they complain and make you feel like you're doing something wrong along the way.

[Walk through an example of saving an Excel file to CSV]

An important note for backwards compatibility: you can open CSVs in Excel!

A cautionary note on Excel for Mac- apparently, there are some issues that can occur in certain versions when you attempt to save to .csv on a Mac- specifically, the line endings generate problems for several statistical problems and coding environments. The GUI-based workaround is to save the file as a "Windows comma separated (.csv)" file. A code-based work-around can be found here:
http://nicercode.github.io/blog/2013-04-30-excel-and-line-endings/