Data and Donuts
Data organization in spreadsheets
Notes

Exercise:
Demo:

Slide 1: Hi, and welcome to Data and Donuts. I'm Tobin Magle, the Cyberinfrastructure Facilitator at the Morgan Library at Colorado State University. Today we're going to talk about best practices for using spreadsheets for data organization.

Slide 2: To put this into the context of the research data lifecycle, data organization is initially important before and during the collection phase, but proper data formatting will also make analysis easier.

Slide 3:
In this first lesson, we'll cover the following data topics:

- How to format data tables in spreadsheets, including common formatting mistakes and formatting dates
- Quality control using the data validation, sorting and conditional formatting features in Excel
- Exporting data from spreadsheets

Slide 4: First, let's talk about how spreadsheets are typically used: like a lab notebook.

- In addition to data, these spreadsheets contain color coding and formatting that often contain information about the data.
- They also have notes, calculations, graphs and tables.
- This is a very human readable and intuitive way to use a spreadsheet.

Slide 5: However, this type of organization has its downsides.

- For one, computers are dumb, so they won't understand the meaning of notes or color coding or formatting
- Also, making graphs and tables in spreadsheet programs is inefficient and hard to reproduce
- Thus, taking the time to format your data in a machine-readable way and learning how to automate your work will save you time in the long run.

Slide 6: Exercise:

1. Put up a green sticky note if you have used spreadsheets in your research.
2. What do you use spreadsheets for? Common answers:
    - Data entry
    - Organizing data
    - Subsetting and sorting data
    - Statistics
    - Plotting

3. Put up your red sticky note if you have ever done something to your spreadsheet data that has made you frustrated or sad. Anyone want to share?

This lesson will show you how to use spreadsheets widely and avoid common pitfalls.

Slide 7:
Now let's talk about how we should be using spreadsheets: for data

1. Spreadsheets are meant for data entry and cleaning
2. But less optimal for calculating statistics, creating summary tables and plotting
3. Example: Generating summary tables: not meant to be read as data (non-machine readable) – heavily formatted – do it in word
4. Example: Statistics and figures – hard to replicate and easy to make mistakes – do it in a statistical package

Slide 8: **And if we want to automate analyses, we have to put our data into a format machines can understand**

1. Computers don't use human readable things like spacing, colors and marginal notes to convey information.
2. To use the power of computers to analyze and visualize data, we need to format data in ways that they understand.

Slide 9:
Machines can read tidy data

1. Leave the raw data raw so you can always go back to the original. This can be accomplished in spreadsheets by creating a new tab.
2. Columns represent variables, or a single piece of information that you're measuring. Make sure not to combine multiple pieces of information in one column!
3. Rows represent observations. Every time you collect a data point, make a new row and add information about each variable in the appropriate column
4. Finally, export cleaned data to CSV, to be more easily readable by statistical programs and preserved for data sharing.

Slide 10:
For example, the data we're using today is from an ecological survey of small animals in a desert ecosystem.

- They are collecting variables like recorded species, plot, weight, sex and date
- For each animal observed, they should make a new row and collect the above variables
- However, different field technicians recorded the data differently, creating inconsistencies.

Slide 11: Exercise:
Why is this table not tidy? How can you fix it?

Slide 12: Species and sex are in same column: separate them into 2 columns

Slide 13: Let's quickly go over some other common data formatting problems

- Multiple tables: computer gets confused: combines multiple observations
- Multiple tabs: can create inconsistencies, creates extra step before data analysis
- Bad null values: for missing values
- Using formatting to convey info: create an extra column
- Bad variable names: not machines readable underscores and camel case

Slide 14: If you consider the fact that computers only really understand "tidy" data, using multiple tables in one sheet can really confuse a computer

- Columns- duplicate column names, empty columns.
- Rows – creates false associations (row = observation)

Slide 15: Using separate data tabs can also be problematic

- Can dissociate things that should be associated: ex – separate tabs for measuring the same things on different days
- Easier to introduce inconsistencies
- Adds a combining step to data analysis

Slide 16: Instead, use tabs wisely:

- New tab for cleaned data
- New tab for analysis steps

Slide 17: It's also important to pick a good symbol for null values

- **Avoid numbers** (especially) because it's impossible to tell whether the value is actually that number or a NULL.
- **Avoid special symbols and uncommon strings** like "missing" because this can cause trouble with data types. For example, importing a number column into R would read in the column as text, not numbers.
- While strings like **NA or NULL** can cause the same data type problems I just mentioned, they are commonly used. Check what your stats program uses
- Overall, **blanks** (not spaces) are the best choice because they won't cause data type problems. One issue is that it's harder to tell if you forgot to enter the data or if you truly have a missing value.

Slide 18: Using colors to indicate data can be problematic as well

- Remember: computers are dumb. They won't understand what a yellow highlighted cell is, and .csvs strip formatting anyway
- Additionally, if you're going to use highlighting to convey meaning, it's good to indicate what it means, as shown here. However, adding notes alongside the table makes the date untidy.
- Exercise: How can we include this information without color?

Slide 19: Solution: make a new column

Slide 20: Also, the way you format your column headings can make or break you

- **Avoid spaces and special characters**
  - most analysis software won't recognize these.
  - Instead: use underscores or camelCase to improve human readability
- **Avoid abbreviations**
  - New data users will have a hard time understanding what variables are
  - Whether you use abbreviations or not, include a codebook that explains what each variable is.

Slide 21: Exercise:

- Download and open the data: http://bit.ly/2dhIsfM
- Make a clean data tab.
- Look at the 2013 and 2014 tabs
- Get a partner: Discuss how to clean up the 2013 and 2014 tabs and combine them into one spreadsheet.

Solution:

- Make one table

Slide 22: Spreadsheet programs can do some pretty funny things with dates, and the behavior depends on what program you're using and even what version of a program.

- So best practice is not to use a date format all in one cell. Instead
- Store dates in 3 columns: Year, Month and Day
- Each one of these columns will store an integer
- If you do want the date all in one column for a table, you can reconstruct with the DATE() function

Slide 23:
Exercise:

- In the dates tab of your spreadsheet you have the data from 2014 plot 3. There's a Date collected column.
- Let's extract month, day and year from the dates to new columns. For this we can use the built in Excel functions MONTH() DAY() YEAR()
- Make sure the new column is formatted as a number and not as a date.

Slide 24: Excel has some features that help ensure data quality.
Let's look at Data Validation first.

- Data validation can stop bad data from being entered and point out errors is previously entered data
- Let's use a semi-cleaned dataset for this: http://bit.ly/2yRKufJ
- Demo: Data ribbon, data validation

Slide 25:
Exercise:

- Select plot_id column
- Enter validation options on right
- What happens when you change the first entry to ’30’?

Solution:
- Data tab>Data Validation
- Allow>Whole Number
- Minimum:1
- Maximum: 24
- Type in 30 -> Value is not valid error message

Slide 26: You can customize the data input messages and errors
- Go back into data validation, change error alert
    - Invalid plot number
    - Must be a number from 1-24
- Enter 30 again
- Go back into data validation, change Input message
    - plot number
    - Enter a number from 1-24
- Note message

Slide 27: You can create a drop down for data entry using the list validation criterion
Demo:
- Make a new tab: metadata, add a header called calibrated scale, values yes, no
- Select Calibrarted_Scale column
- Data > Data validation
- Settings: Allow list
- Source: Yes or no from the metadata tab
- Show dropdown

Slide 28: You can also inspect your data by sorting
Demo:
- Data >Sort
- Look at top and bottom to see what is out of range
- Example: weight_grams
- What do you notice?
- Solution: ##g gets sorted out

Slide 29: Finally, when your data is in good shape, you can export it to CSV:
- File, save as
- File format: Comma Separated Values (.csv)

Slide 30: Exercise: look at some other data. Tidy it up
- Lou’s data
- Two real life examples from FigShare

Slide 31:

Download Lou's files: http://bit.ly/2BrnOVv

How can you tidy up his mouse inventory?

- Separate variables into columns
- Add a column for sick mouse formatting

How can you combine all weight data into one spreadsheet? What about the cytokine data?

- Create a new sheet
- Add a column for month
- Paste rows beneath.

Slide 32:

Now let's look at some examples of real data posted to FigShare, which houses real shared data. The first is supplemental data from a manuscript. This record contains high performance liquid chromatography data, which separates out molecules in solution to get an idea of how big they are. Primarily, they are measuring retention time, which is how long a molecule stays on the column. Let's look at how these data are organized.

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

- Rt = Retention time
- + = presence of a peak
- - = absence of a peak
- **Variables**: Sample Code, Species, Sampling area, Altitude, Lat, Lon, Retention time
- **Observation**, whether there is a peak (+/-) at the combination of the variables above
- Lots of "extra" data. Could assume that no data = absence of a peak.
- Make new tab called tidy
- Put variables across the top
    - Copy A1 – A5
    - Paste transpose
    - Add retention time
- Instead of having a series of +/- for each combo of variables and each retention time, make a row for where each + sign is
    - Each combo of variables can have more than one record.
    - Hard to do by hand: scan for + , add records manually
    - Easier to do in R.
    - Results in way fewer data points -> efficient storage.

Slide 33:

This next dataset contains cytokine data from a CCK-8 assay and ELISA. It's hard to say what all of these data mean, because the metadata are insufficient, but we'll do our best.

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

- First, you can see they have multiple sheets, but sheets 2 and 3 are empty, so they can be deleted
- It also looks like they have 2 separate assays (CCK-8 and ELISA). I would make a separate sheet for each assay.
- Let's start with the CCK8 data:
    - **Variables**: Concentration (uC/ml), time, treatment, value
    - **Observations**: value at a certain treatment, concentration and time
- For ELISA:
    - **Variables**: Cytokine, treatment, value
    - **Observation**: level of a specified cytokine given the treatment

Slide 34: Thanks for listening. I hope you found this data organization session to be helpful. Please email me at the address on this slide if you need help. Also, check out our data management pages for more information. If you want to learn more about data organization, check out the Ecology spreadsheet lessons from Data carpentry.