

Data for Data Science

fredhutch.io

Fred Hutchinson Cancer Research Center

Sign in here: <https://hackmd.io/@k8hertweck/introData>

Motivation: Why learn about data management?

- Scientific research is increasingly data-driven
- Few researchers are adequately trained to organize, manage, and track scientific data
- An understanding of best practices in data management can help save you time, energy, and frustration later

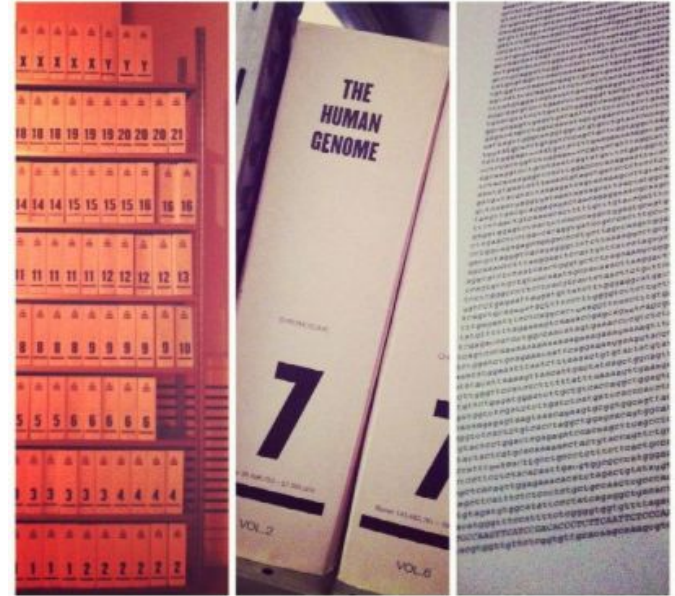


Photo by Erin Morris

Course outline

- **Class 1: Data entry and creating spreadsheets**
- Class 2: Organizing data and project files
- Class 3: Documenting data with metadata
- Class 4: Data manipulation and reproducibility

At the end of this course, you should be able to apply *some* best practices in data management to your *current* research projects...

...and *plan ahead* to improve your *future* projects

Today's objectives

At the end of today's class, you should be able to:

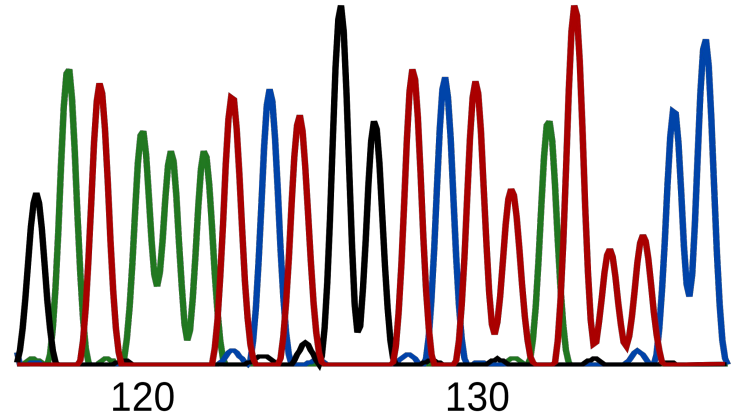
- Implement best practices in data table formatting
- Identify and address common formatting mistakes
- Understand approaches for handling dates in spreadsheets
- Effectively export data from spreadsheet programs

Understanding best practices for data entry and formatting is foundational to other issues in data management

Why spreadsheets?

Biological data is often represented as text

Data as text can be interpreted by both humans and computers, although not always as easily for both



GAT AAATCT GGTCCTTATTTC

Why spreadsheets?

Spreadsheets provide a useful, familiar structure for working with tabular data (organized in rows and columns)

Spreadsheet programs like Microsoft Excel LibreOffice/OpenOffice allow us to view these data in a comprehensible way

What are some limitations of spreadsheets? Have you ever had a frustrating experience with a spreadsheet program?

Organizing existing data in spreadsheets

Spreadsheets are a great way to enter and organize data, some of their features and the habits we develop through using them can make it difficult to perform data science tasks with these same data later.

Our goal is to have data entered in a way that we can easily export in a plain-text format that is straightforward for other programs to interpret

experiment	sex	weight_g	treatment
4	F	41	control
3	F	37	control
4	F	117	insulin
3	F	121	insulin
2	M	115	insulin



```
experiment,sex,weight_g,treatment
4,F,41,control
3,F,37,control
4,F,117,insulin
3,F,121,insulin
2,M,115,insulin
```

Tidy data

Each cell should represent a single **value** (piece of data)

Rows represent individual **observations** (samples, patients, etc)

Columns represent **variables** (information about each observation)

The diagram shows a table with 5 rows and 4 columns. Above the table, the word 'variables' is centered, with four arrows pointing down to each of the four column headers. To the left of the table, the word 'observation' is followed by an arrow pointing to the first row. To the right of the table, the word 'value' is preceded by an arrow pointing to the cell containing 'control' in the second row.

experiment	sex	weight_g	treatment
4	F	41	control
3	F	37	control
4	F	117	insulin
3	F	121	insulin
2	M	115	insulin

Organizing existing data in spreadsheets

If you have data that need to be organized, keep the following guidelines in mind:

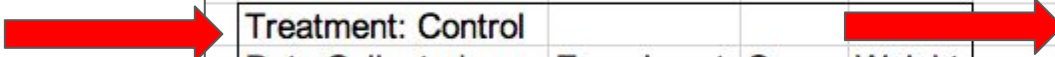
- Keep the original data unaltered
- Create a new spreadsheet to hold the reformatted data
- List the steps taken while reformatting in another spreadsheet
- Export the final (reformatted) spreadsheet as a text file

Spreadsheet activity:

Using `experimental_data.xlsx`, put the data from the 2013 and 2014 tabs into the same table.

Common spreadsheet errors

Multiple tables appearing in the same sheet




2013 Experiments						
Treatment: Control					Treatment: Insulin	
Date Collected	Experiment	Sex	Weight		Date Collected	Ex
7/16/13	1	F			11/12/13	
7/16/13	2	M	33g		11/12/13	
7/16/13	4	M			11/12/13	
7/16/13	3	M			11/12/13	
7/18/13	4	M	40g		11/13/13	
7/18/13	2	M	48g		11/13/13	

Common spreadsheet errors

Related data split among multiple sheets/tabs

7/18/13	2 M	36g			
7/18/13	1 F	35g			
7/18/13	2 F	22g			
7/18/13	3 F	42g			
7/18/13	4 F	41g			
7/18/13	3 F	37g			



2013

2014

dates

+

Common spreadsheet errors

Inconsistent use of zeros versus missing data (null/blank values)

Zero (0) can represent observed data

Does a blank cell indicate no data was collected, or an error in data entry?

If all cells must have data, what symbol represents missing values?

Experiment: 3			
Date collected	Species	Sex	Weight
1/8	Control	M	7
2/18	Insulin	M	24
2/18	Control	F	23
3/11	Insulin	M	232
3/11	Control	F	22
3/11	Control	M	26
3/11	Insulin	M	8
4/8	Insulin	F	0
5/6			
5/18	Control	F	182
6/9	Insulin	F	29
7/8	Control	F	115
7/8	Insulin	M	190

Common spreadsheet errors




Using formatting to convey information

	1/8/14	Control						
	1/8/14	Insulin	M	157				
	1/8/14	insulin						
	2/18/14	control	M	218				
	2/18/14	Insulin	F	7				
	2/18/14	Insulin	M	52				
		gray cell means my measurement device wasn't calibrated correctly						

Common spreadsheet errors

Formatting the spreadsheet to make the table easier to read

- Merging and splitting cells
- Including empty rows or columns



B		C	D	E	F
Treatment: Control					
Date Collected		Experiment		Sex	Weight
7/16/13		1		F	
7/16/13		2		M	33g
7/16/13		4		M	

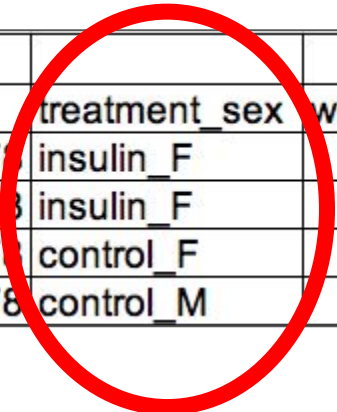
Common spreadsheet errors

Including comments or units in cells

B		C	D	E	F
Treatment: Control					
Date Collected		Experiment		Sex	Weight
7/16/13		1		F	
7/16/13		2		M	33g
7/16/13		4		M	
7/16/13					

Common spreadsheet errors

Including more than one piece of data in a single cell



Experiment: 4		
Date collected	treatment_sex	wgt
1/8/78	insulin_F	37
1/8/78	insulin_F	128
1/8/78	control_F	42
1/8/78	control_M	37

Common spreadsheet errors

Using special characters in data tables, as these can have unintended consequences during data analysis later

! " ' SPACE # \$ % & ' () * + , - . / : ; < = > ? @ [\] ^ ` { | } ~

New lines, tabs, and vertical tabs

Common spreadsheet errors

Problematic field names (row and column labels)

Avoid spaces: use underscores (or dashes, or capitalization)

Good Name	Good Alternative	Avoid
Max_temp_C	MaxTemp	Maximum Temp (°C)
Precipitation_mm	Precipitation	precmm
Mean_year_growth	MeanYearGrowth	Mean growth/year
sex	sex	M/F
weight	weight	w.
cell_type	CellType	Cell Type
Observation_01	first_observation	1st Obs

Common spreadsheet errors

Including metadata in a data file

Metadata (data about data, e.g., descriptions of what variables represent) should go in a separate file

date	experiment	sex	weight_g	treatment	calibrated
7/16/13		1 F	NA	control	yes
7/16/13		2 M	33	control	yes
7/16/13		4 M	NA	control	yes
7/16/13		3 M	NA	control	yes
date	month/day/year				
experiment	1 through 4				
sex	M (male) or F (female)				
weight_g	weight in grams				
treatment	control or insulin				
calibrated	was scale calibrated appropriately?				
Data collected by Research Technican X in Professor Y's Lab					

Common spreadsheet errors

Unclear formatting of dates as data

Spreadsheet programs include features that allow manipulation of dates

	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

These features can accidentally introduce ambiguity or conversion errors

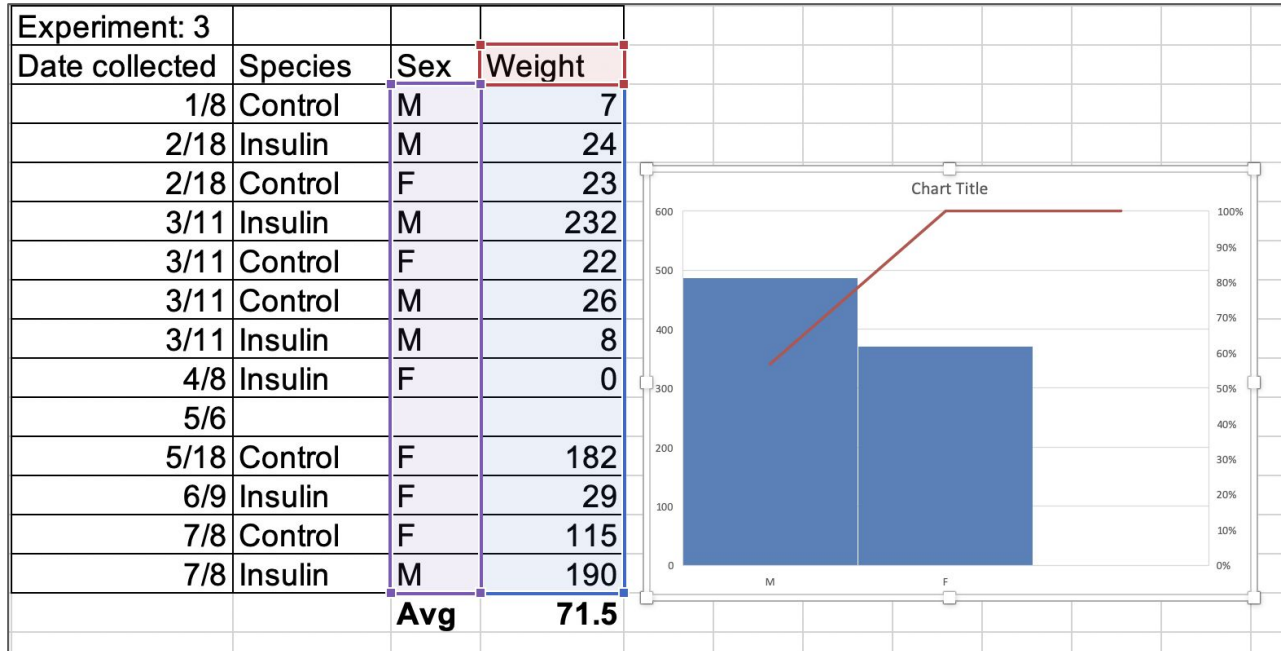
Recommendation: store year, month, and day as separate columns



® The International Organization for Standardization (ISO) also has a format ([ISO 8601](#)) for dates and time

Common spreadsheet errors

Including calculations and/or figures in data file



Exporting data as a text file

Spreadsheet programs do a lot more than help us organize data, and those things may interfere with long-term data stability and interpretation.

Export your data from a spreadsheet program to plain text file, which is universal, open, and static (meaning more stable and accessible)

experiment	sex	weight_g	treatment
4	F	41	control
3	F	37	control
4	F	117	insulin
3	F	121	insulin
2	M	115	insulin



```
experiment,sex,weight_g,treatment
4,F,41,control
3,F,37,control
4,F,117,insulin
3,F,121,insulin
2,M,115,insulin
```

Exporting data as a text file

You can save your data as a text file using the “Save as” command in a spreadsheet program

Common formats are .tsv (tab separated values) and .csv (comma separated values); they are differentiated by the character that separates the columns on each line (there are other options as well, but these are most common)

csv

```
experiment,sex,weight_g,treatment
4,F,41,control
3,F,37,control
4,F,117,insulin
3,F,121,insulin
2,M,115,insulin
```

tsv

```
experiment sex  weight_g
treatment
4      F      41      control
3      F      37      control
4      F      117     insulin
3      F      121     insulin
2      M      115     insulin
```

These files can still be opened by spreadsheet programs!

A bigger picture of the data lifecycle



Data is valuable, and often includes the interest of many stakeholders.

It may not be possible to incorporate best practices for long-running projects that are nearing completion.

Understanding best practices will help you identify which areas you can improve, and how to plan better for the future.

Summary

- Spreadsheets are a great way to enter and organize data, and can be used to apply tidy data principles
- Take care to avoid common errors associated with data organization and formatting in spreadsheets to make your life easier when it's time to analyze data
- Saving your data files as .csv or .tsv makes your data more stable and accessible

Next time: organizing data files

Resources

- These materials were adapted from [this lesson](#) from [Data Carpentry](#)
- [Tidy Data](#) by Hadley Wickham describes these principles in more detail
- The course website is publicly available [here](#)
- [Fred Hutch Biomedical Data Science Wiki](#) contains information on resources at Fred Hutch for work with data and computational analysis