# Introduction to Research Data Management

Data collection

http://hdl.handle.net/1969.1/164246

LIBRARIES
TEXAS A&M UNIVERSITY

# Workshops

1. Build an overview

**2. Collect and document data**

3. Store digital data

4. Work with data

5. Share and preserve data

6. Plan ahead

# Introduction

Focus on

- gathering data,
- data entry,
- and arranging data to your advantage.

# Ways to gather data

- Observational data
- Experimental data
- Simulation data
- Derived and Compiled data
- Reference data

# Discussion

Which source(s) of data do you have experience using, or expect to use, in your graduate research?

Follow up: If these data are lost, what are potential problems or limitations with recreating data from the source(s) you listed?

# Reproducibility

**Observational**: Usually irreplaceable and important to safeguard.

**Experimental**: Often reproducible, can be expensive or time-consuming.

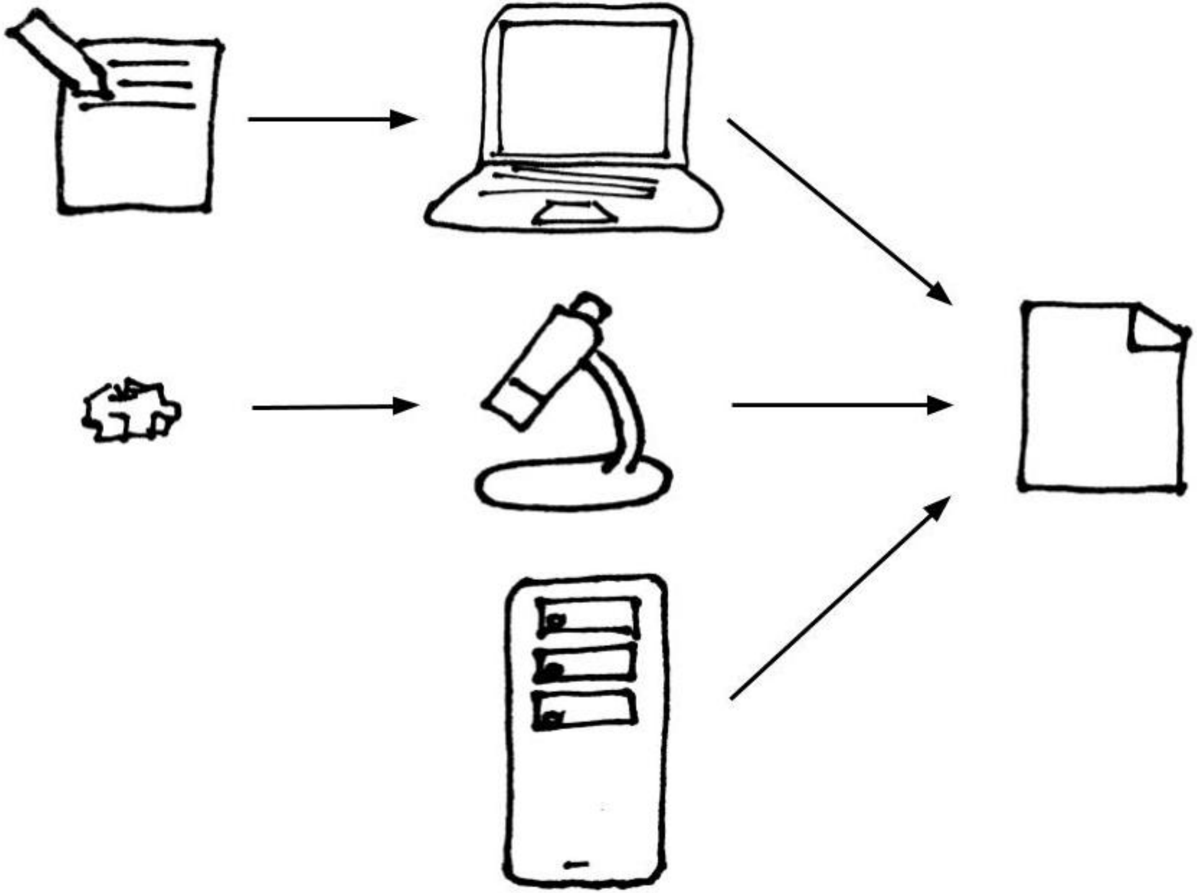
**Simulation**: Likely to be reproducible, if the model and inputs are preserved.

**Derived and compiled data**: Reproducible, but can be very expensive and time-consuming.

# Video: Data loss

Bruce Herbert, Professor of Geology at Texas A&M University.

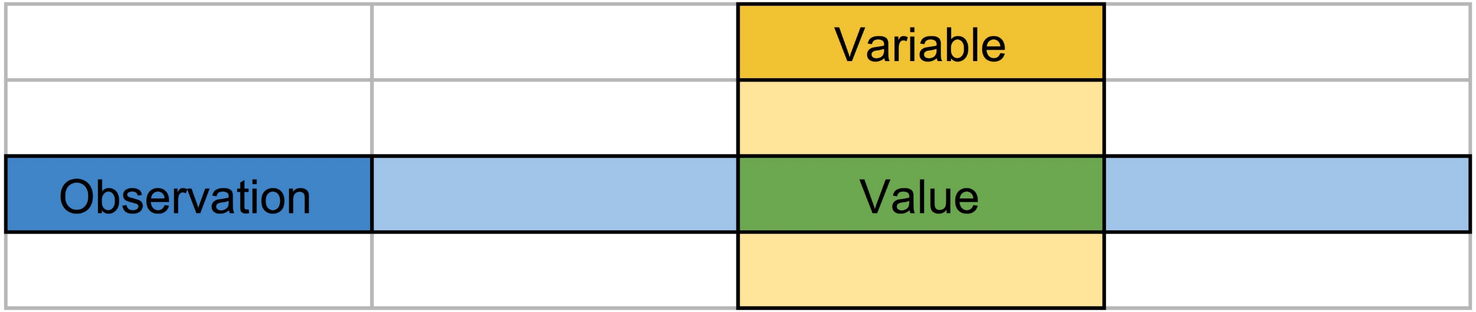
https://youtu.be/xy7b_6MIB4k



# Tips to assure quality

- Enforce consistent procedure and use of standards.
- Minimize number of times data must be entered.
- Consider adding an automatic validation layer.
- Double check for errors. Preferably using a second person.

# Tabular data structure



| | | | |
|---|---|---|---|
| | | Variable | |
| | | | |
| Observation | | Value | |
| | | | |

# Tabular data structure

OPEN Refine
R
SPSS
sas
MATLAB

# Exercise: Common errors in spreadsheets

How many different errors can you identify?

Why are they problematic, and what changes should be made?

Data file: https://ndownloader.figshare.com/files/2252083

# 6 tips to avoid problems

1. Add column names in the first row(s), or "header".

2. Use descriptive column names without spaces or special characters.

3. Include one observation per row, and one variable per column.

4. Ensure that data in each column are of a single type.

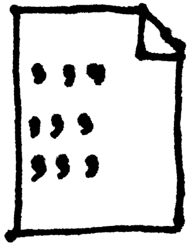
5. Use a standard format and code(s) for values in each column.

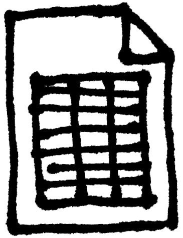
6. Once data are captured, leave the raw data raw.

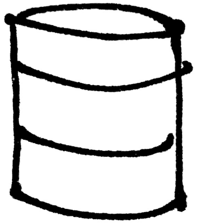
# Using codes

- Identify missing data
- “Dummy codes” for surveys
- Standard codes for discipline-specific values

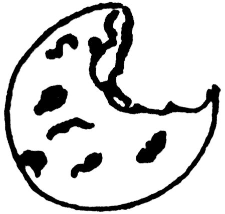
# Formats and tools







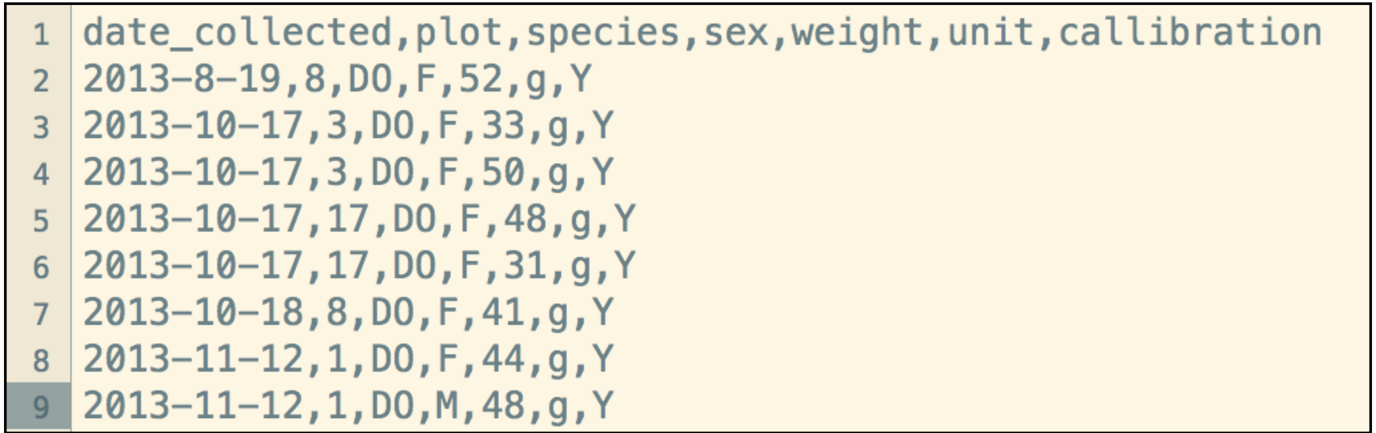
# Break



# “Plain text”

Best for a single tabular dataset, especially when it is simple and large.

Easy automatic data entry and programmatic manipulation.



```
date_collected,plot,species,sex,weight,unit,callibration
2013-8-19,8,DO,F,52,g,Y
2013-10-17,3,DO,F,33,g,Y
2013-10-17,3,DO,F,50,g,Y
2013-10-17,17,DO,F,48,g,Y
2013-10-17,17,DO,F,31,g,Y
2013-10-18,8,DO,F,41,g,Y
2013-11-12,1,DO,F,44,g,Y
2013-11-12,1,DO,M,48,g,Y
```

# “Plain text”

Saved as comma-separated value (CSV) files or tab-separated values files (TSV).

Software-agnostic file format accessible to:

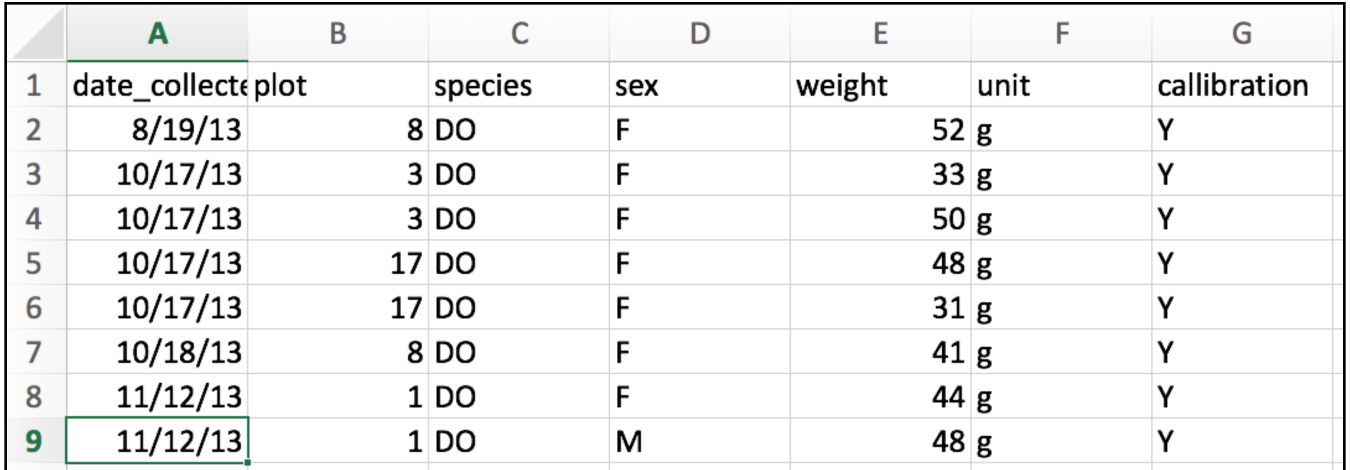
- text editing software
- spreadsheet software
- database software
- statistical programming environments

# Spreadsheets

Best for single or multiple tabular datasets.

Interface to enter and manipulate data manually.

Additional functionality for cleaning, analyzing, and presenting data.



| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | date_collecte | plot | species | sex | weight | unit | callibration |
| 2 | 8/19/13 | 8 | DO | F | 52 | g | Y |
| 3 | 10/17/13 | 3 | DO | F | 33 | g | Y |
| 4 | 10/17/13 | 3 | DO | F | 50 | g | Y |
| 5 | 10/17/13 | 17 | DO | F | 48 | g | Y |
| 6 | 10/17/13 | 17 | DO | F | 31 | g | Y |
| 7 | 10/18/13 | 8 | DO | F | 41 | g | Y |
| 8 | 11/12/13 | 1 | DO | F | 44 | g | Y |
| 9 | 11/12/13 | 1 | DO | M | 48 | g | Y |

# Spreadsheets

Multiple file formats, based on software.

# Databases

Work best for linking large and complex datasets.

Easy to query datasets, select portions and combine.

Steep learning curve and rigid in use.

Control errors by forced typing and maintaining record integrity.

# Databases

A database consists of

1. a set of tables,

2. defined relationships between tables,

3. a command language that facilitates data manipulation.

WeightData

| date_collected | plot_id | species | sex | weight | unit | callibration |
|---|---|---|---|---|---|---|
| 2013-08-19 | 6 | DO | F | 52 | g | Y |
| 2013-10-17 | 3 | DM | F | 33 | g | Y |
| 2013-10-17 | 3 | DO | F | 50 | g | Y |
| 2013-10-17 | 4 | DS | F | 48 | g | Y |
| 2013-10-17 | 4 | DO | F | 31 | g | Y |
| 2013-10-18 | 6 | DS | F | 41 | g | Y |
| 2013-11-12 | 1 | DO | F | 44 | g | Y |
| 2013-11-12 | 1 | DM | M | 48 | g | Y |

Plot

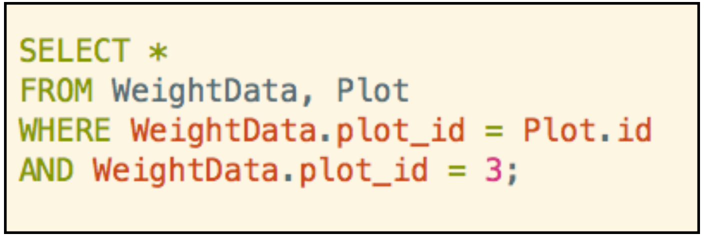
| id | city | state | country_code |
|---|---|---|---|
| 1 | College Station | Texas | USA |
| 2 | Bryan | Texas | USA |
| 3 | San Antonio | Texas | USA |
| 4 | Bastrop | Texas | USA |
| 5 | Huntsville | Texas | USA |
| 6 | Stephenville | Texas | USA |

WeightData

| date_collected | plot_id | species | sex | weight | unit | callibration |
|---|---|---|---|---|---|---|
| 2013-08-19 | 6 | DO | F | 52 | g | Y |
| 2013-10-17 | 3 | DM | F | 33 | g | Y |
| 2013-10-17 | 3 | DO | F | 50 | g | Y |
| 2013-10-17 | 4 | DS | F | 48 | g | Y |
| 2013-10-17 | 4 | DO | F | 31 | g | Y |
| 2013-10-18 | 6 | DS | F | 41 | g | Y |
| 2013-11-12 | 1 | DO | F | 44 | g | Y |
| 2013-11-12 | 1 | DM | M | 48 | g | Y |

Plot

| id | city | state | country_code |
|---|---|---|---|
| 1 | College Station | Texas | USA |
| 2 | Bryan | Texas | USA |
| 3 | San Antonio | Texas | USA |
| 4 | Bastrop | Texas | USA |
| 5 | Huntsville | Texas | USA |
| 6 | Stephenville | Texas | USA |



```
SELECT *
FROM WeightData, Plot
WHERE WeightData.plot_id = Plot.id
AND WeightData.plot_id = 3;
```

WeightData

| date_collected | plot_id | species | sex | weight | unit | callibration |
|---|---|---|---|---|---|---|
| 2013-08-19 | 6 | DO | F | 52 | g | Y |
| 2013-10-17 | 3 | DM | F | 33 | g | Y |
| 2013-10-17 | 3 | DO | F | 50 | g | Y |
| 2013-10-17 | 4 | DS | F | 48 | g | Y |
| 2013-10-17 | 4 | DO | F | 31 | g | Y |
| 2013-10-18 | 6 | DS | F | 41 | g | Y |
| 2013-11-12 | 1 | DO | F | 44 | g | Y |
| 2013-11-12 | 1 | DM | M | 48 | g | Y |

Plot

| id | city | state | country_code |
|---|---|---|---|
| 1 | College Station | Texas | USA |
| 2 | Bryan | Texas | USA |
| 3 | San Antonio | Texas | USA |
| 4 | Bastrop | Texas | USA |
| 5 | Huntsville | Texas | USA |
| 6 | Stephenville | Texas | USA |

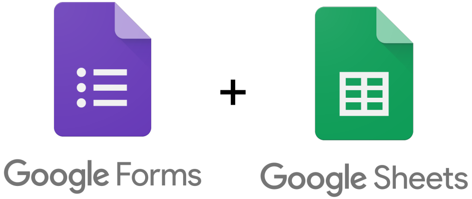
```
SELECT *
FROM WeightData, Plot
WHERE WeightData.plot_id = Plot.id
AND WeightData.plot_id = 3;
```

| date_collected | plot_id | species | sex | weight | unit | callibration | city | state | country_code | state | country_code |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2013-10-17 | 3 | DM | F | 33 | g | Y | San Antonio | Texas | USA | Texas | USA |
| 2013-10-17 | 3 | DO | F | 50 | g | Y | San Antonio | Texas | USA | Texas | USA |

# Databases

Multiple file formats and models, based on software.

# Using automatic validation



# Conclusion

- Discussed data sources and reproducibility.
- Compared tabular data arrangements in spreadsheets.
- Reviewed differences in plain text, spreadsheet, and database formats.

# References and resources

- Data Carpentry. "Data Organization in Spreadsheets" [Website](http://www.datacarpentry.org/spreadsheet-ecology-lesson/)
- Cornell University. “Preparing tabular data for description and archiving” [Website](https://data.research.cornell.edu/content/tabular-data)
- DataONE. "Data entry and manipulation" [Website](https://www.dataone.org/education-modules)
- DMPTool. "Data Management General Guidance" [Website](https://dmptool.org/dm_guidance)
- Litwin, Paul. “Fundamentals of Relational Database Design” [Website](http://r937.com/relational.html)
- Wickham, Hadley. 2014. Tidy Data. Journal of Statistical Software, 59:10 [Article](https://doi.org/10.18637/jss.v059.i10)