Data Entry and Manipulation

GEO 802 Fall 2020, Data Information Literacy

Anna C. Véron, Dr. sc. nat.

Learning Objectives

- Recognize inconsistencies that can make a **dataset difficult to understand** and/or manipulate
- Identify **data entry tools**
- Identify **validation measures** that can be performed as data is entered
- Describe the basic components of a **relational database**

→ **The best way to record your data varies from discipline to discipline.**

→ **You decide what is best for your data!**

Lesson 3: Data Entry and Manipulation

→How to structure your data: Best practices

- Quality of research data
- Data entry tools
- Databases
- Data Analysis

Collecting data: everyone does it a «little different»? – Better not!

CC image by Travis S on Flickr

Create datasets that are **valid** and **structured**.

Enter your data into **spreadsheets** or a **database**, especially when **collaboratively** working on a dataset.

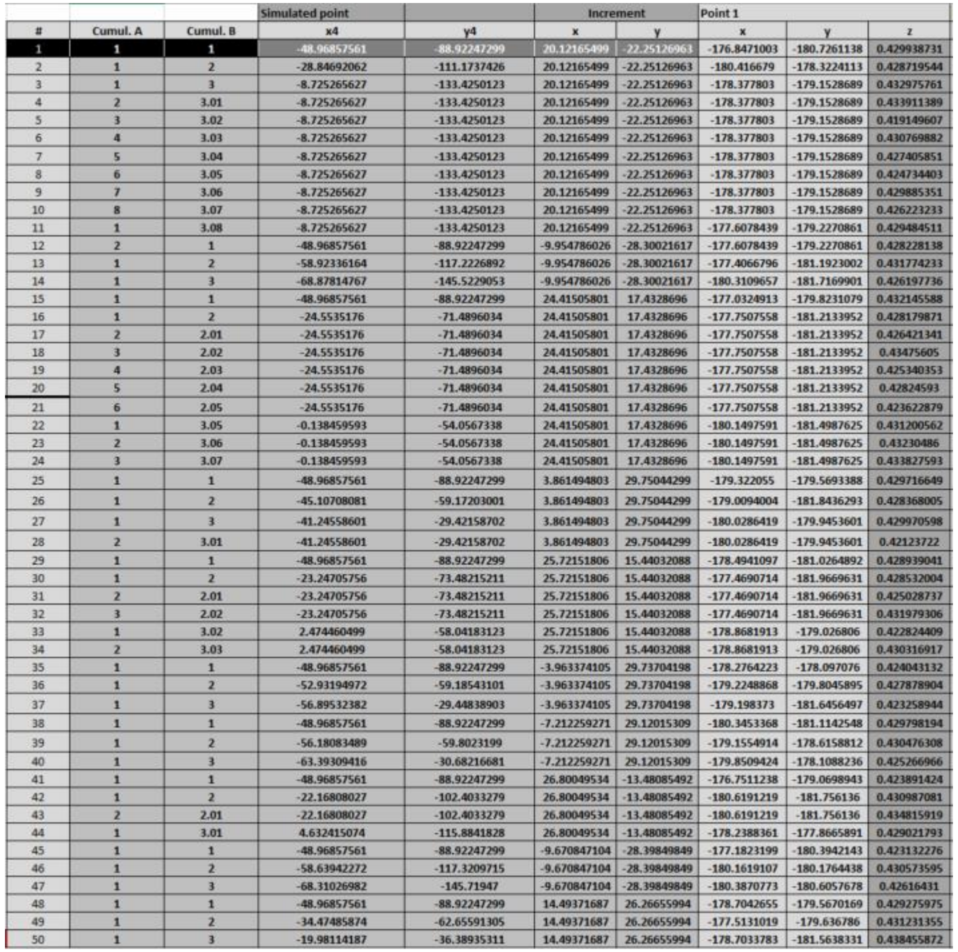
Structured vs. Unstructured data

Structured data

- Highly organized, usually text-only
- Pre-defined data models
- easy to access, search and analyze scientifically
- (usually) machine-readable

Sources of Structured Data:

- SQL databases, spreadsheets, XML, tables
- Sensors, measurement instruments
- Medical devices
- Online forms
- **People who enter data into spreadsheets and databases**
- etc.



			Simulated point		Increment		Point 1		
#	Cumul. A	Cumul. B	x4	y4	x	y	x	y	z
1	1	1	-48.96857561	-88.92247299	20.12165499	-22.25126963	-176.8471003	-180.7261138	0.429938731
2	1	2	-28.84692062	-111.1737426	20.12165499	-22.25126963	-180.416679	-178.3224113	0.428719544
3	1	3	-8.725265627	-133.4250123	20.12165499	-22.25126963	-178.377803	-179.1528689	0.432975761
4	2	3.01	-8.725265627	-133.4250123	20.12165499	-22.25126963	-178.377803	-179.1528689	0.433911389
5	3	3.02	-8.725265627	-133.4250123	20.12165499	-22.25126963	-178.377803	-179.1528689	0.419149607
6	4	3.03	-8.725265627	-133.4250123	20.12165499	-22.25126963	-178.377803	-179.1528689	0.430769882
7	5	3.04	-8.725265627	-133.4250123	20.12165499	-22.25126963	-178.377803	-179.1528689	0.427405851
8	6	3.05	-8.725265627	-133.4250123	20.12165499	-22.25126963	-178.377803	-179.1528689	0.424734403
9	7	3.06	-8.725265627	-133.4250123	20.12165499	-22.25126963	-178.377803	-179.1528689	0.429885351
10	8	3.07	-8.725265627	-133.4250123	20.12165499	-22.25126963	-178.377803	-179.1528689	0.426223233
11	1	3.08	-8.725265627	-133.4250123	20.12165499	-22.25126963	-177.6078439	-179.2270861	0.429484511
12	2	1	-48.96857561	-88.92247299	-9.954786026	-28.30021617	-177.6078439	-179.2270861	0.428228138
13	1	2	-58.92336164	-117.2226892	-9.954786026	-28.30021617	-177.4066796	-181.1923002	0.431774233
14	1	3	-68.87814767	-145.5229053	-9.954786026	-28.30021617	-180.3109657	-181.7169901	0.426197736
15	1	1	-48.96857561	-88.92247299	24.41505801	17.4328696	-177.0324913	-179.8231079	0.432145588
16	1	2	-24.5535176	-71.4896034	24.41505801	17.4328696	-177.7507558	-181.2133952	0.428179871
17	2	2.01	-24.5535176	-71.4896034	24.41505801	17.4328696	-177.7507558	-181.2133952	0.426421341
18	3	2.02	-24.5535176	-71.4896034	24.41505801	17.4328696	-177.7507558	-181.2133952	0.43475605
19	4	2.03	-24.5535176	-71.4896034	24.41505801	17.4328696	-177.7507558	-181.2133952	0.425340353
20	5	2.04	-24.5535176	-71.4896034	24.41505801	17.4328696	-177.7507558	-181.2133952	0.42824593
21	6	2.05	-24.5535176	-71.4896034	24.41505801	17.4328696	-177.7507558	-181.2133952	0.423622879
22	1	3.05	-0.138459593	-54.0567338	24.41505801	17.4328696	-180.1497591	-181.4987625	0.431200562
23	2	3.06	-0.138459593	-54.0567338	24.41505801	17.4328696	-180.1497591	-181.4987625	0.43230486
24	3	3.07	-0.138459593	-54.0567338	24.41505801	17.4328696	-180.1497591	-181.4987625	0.433827593
25	1	1	-48.96857561	-88.92247299	3.861494803	29.75044299	-179.322055	-179.5693388	0.429716649
26	1	2	-45.10708081	-59.17203001	3.861494803	29.75044299	-179.0094004	-181.8436293	0.428368005
27	1	3	-41.24558601	-29.42158702	3.861494803	29.75044299	-180.0286419	-179.9453601	0.429970598
28	2	3.01	-41.24558601	-29.42158702	3.861494803	29.75044299	-180.0286419	-179.9453601	0.42123722
29	1	1	-48.96857561	-88.92247299	25.72151806	15.44032088	-178.4941097	-181.0264892	0.428939041
30	1	2	-23.24705756	-73.48215211	25.72151806	15.44032088	-177.4690714	-181.9669631	0.428532004
31	2	2.01	-23.24705756	-73.48215211	25.72151806	15.44032088	-177.4690714	-181.9669631	0.425028737
32	3	2.02	-23.24705756	-73.48215211	25.72151806	15.44032088	-177.4690714	-181.9669631	0.431979306
33	1	3.02	2.474460499	-58.04183123	25.72151806	15.44032088	-178.8681913	-179.026806	0.422824409
34	2	3.03	2.474460499	-58.04183123	25.72151806	15.44032088	-178.8681913	-179.026806	0.430316917
35	1	1	-48.96857561	-88.92247299	-3.963374105	29.73704198	-178.2764223	-178.097076	0.424043132
36	1	2	-52.93194972	-59.18543101	-3.963374105	29.73704198	-179.2248868	-179.8045895	0.427878904
37	1	3	-56.89532382	-29.44838903	-3.963374105	29.73704198	-179.198373	-181.6456497	0.423258944
38	1	1	-48.96857561	-88.92247299	-7.212259271	29.12015309	-180.3453368	-181.1142548	0.429798194
39	1	2	-56.18083489	-59.8023199	-7.212259271	29.12015309	-179.1554914	-178.6158812	0.430476308
40	1	3	-63.39309416	-30.68216681	-7.212259271	29.12015309	-179.8509424	-178.1088236	0.425266966
41	1	1	-48.96857561	-88.92247299	26.80049534	-13.48085492	-176.7511238	-179.0698943	0.423891424
42	1	2	-22.16808027	-102.4033279	26.80049534	-13.48085492	-180.6191219	-181.756136	0.430987081
43	2	2.01	-22.16808027	-102.4033279	26.80049534	-13.48085492	-180.6191219	-181.756136	0.434815919
44	1	3.01	4.632415074	-115.8841828	26.80049534	-13.48085492	-178.2388361	-177.8665891	0.429021793
45	1	1	-48.96857561	-88.92247299	-9.670847104	-28.39849849	-177.1823199	-180.3942143	0.423132276
46	1	2	-58.63942272	-117.3209715	-9.670847104	-28.39849849	-180.1619107	-180.1764438	0.430573595
47	1	3	-68.31026982	-145.71947	-9.670847104	-28.39849849	-180.3870773	-180.6057678	0.42616431
48	1	1	-48.96857561	-88.92247299	14.49371687	26.26655994	-178.7042655	-179.5670169	0.429275975
49	1	2	-34.47485874	-62.65591305	14.49371687	26.26655994	-177.5131019	-179.636786	0.431231355
50	1	3	-19.98114187	-36.38935311	14.49371687	26.26655994	-178.7033783	-181.5638331	0.438455872

Borges, C., Palma, C. & da Silva, R. B.. Optimization of River Sampling: Application to Nutrients Distribution in Tagus River Estuary (2019). https://doi.org/10.1021/acs.analchem.8b05781.s001

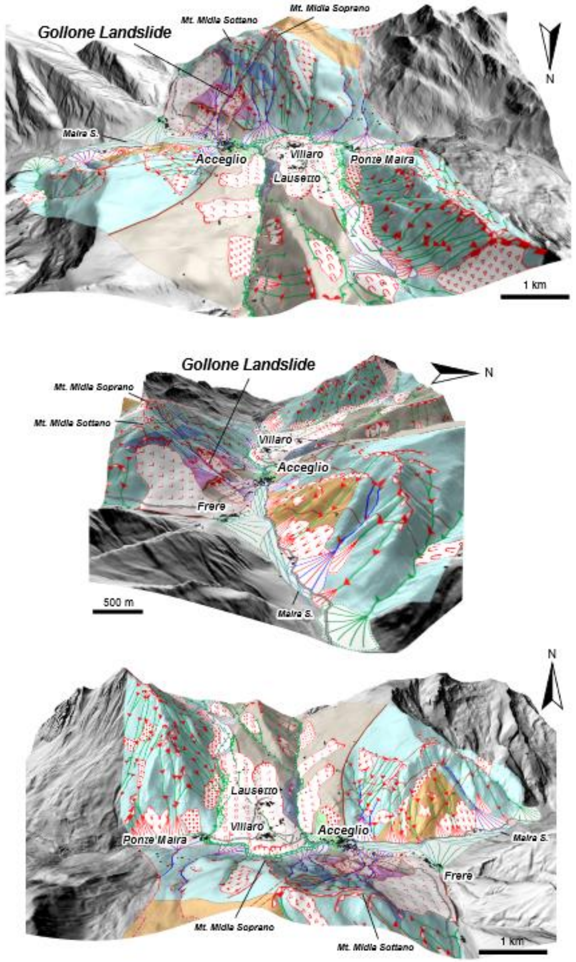
Structured vs. Unstructured data

Unstructured data

- No pre-defined data model
- Difficult to search
- Not «machine-readable», but can be analyzed with text mining, data mining and AI techniques (time-consuming)
- More than 80% of data generated in the world

Sources of Unstructured Data:

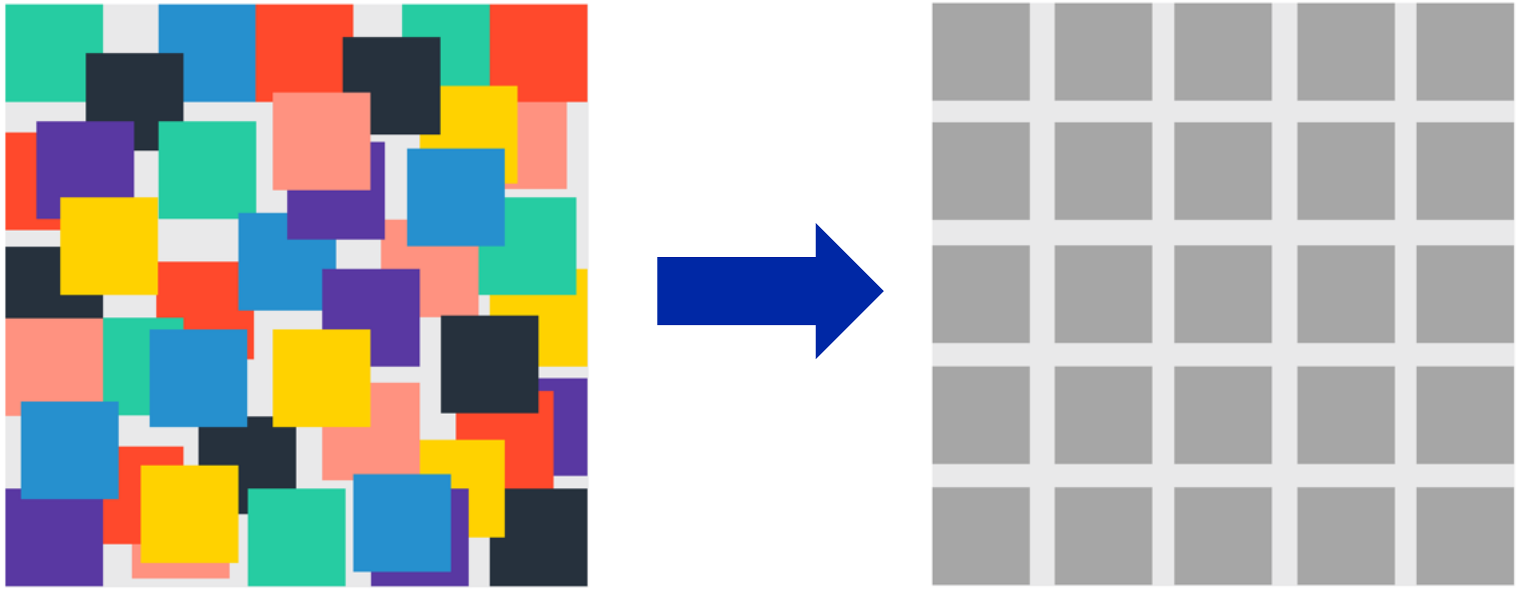
- Text files, presentations, emails, websites, diaries
- Social media, text messages, chat
- image, audio and video files
- **Examples from Science:** satellite imagery, microscope images, space exploration, seismic imagery, atmospheric data, surveillance photos / videos, sensor data



Petroccia, A.. Structural and geomorphological framework of the upper Maira Valley (Western Alps, Italy): the case study of the Gollone Landslide (2020). https://doi.org/10.6084/m9.figshare.12854354.v1

Structured vs. Unstructured data

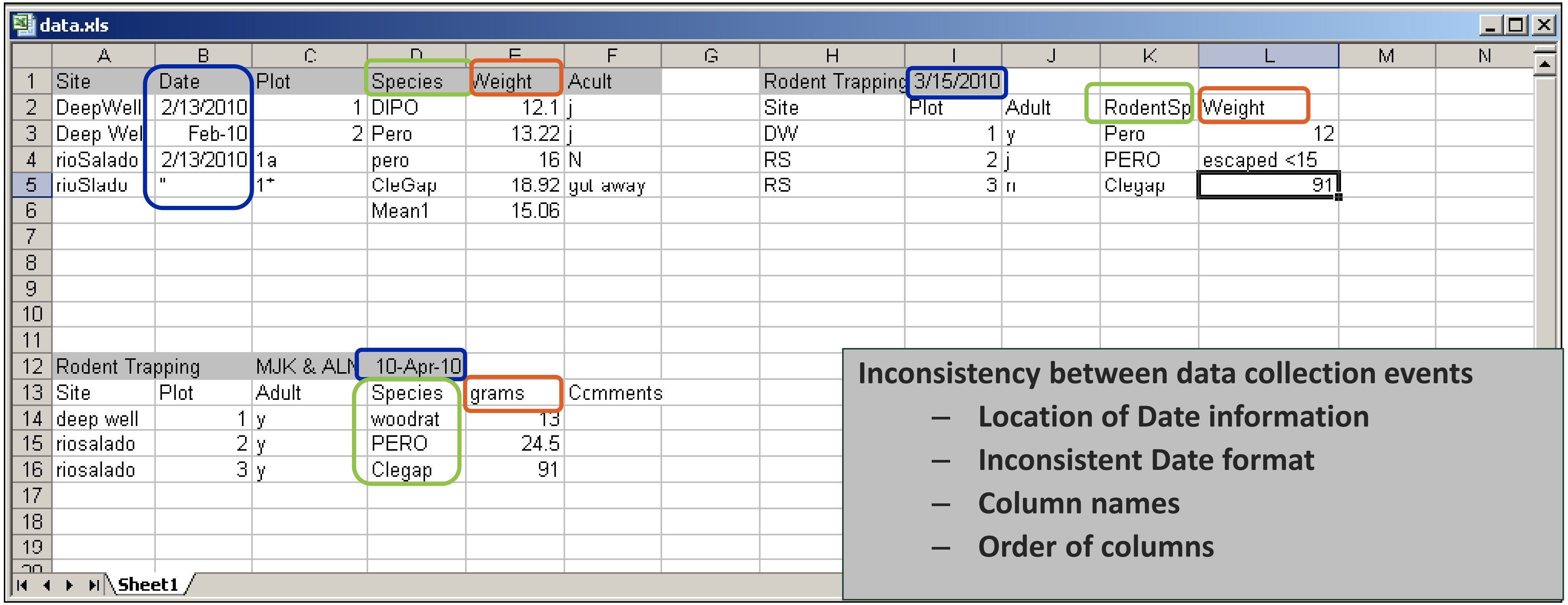
Whenever possible, create structured data!



Devin Pickell, G2 Learning Hub, Structured vs. Unstructured Data – What's the Difference?
https://learn.g2.com/structured-vs-unstructured-data; accessed Aug 26th 2020

Example: unstructured data entry

From a small mammal trapping study



data.xls

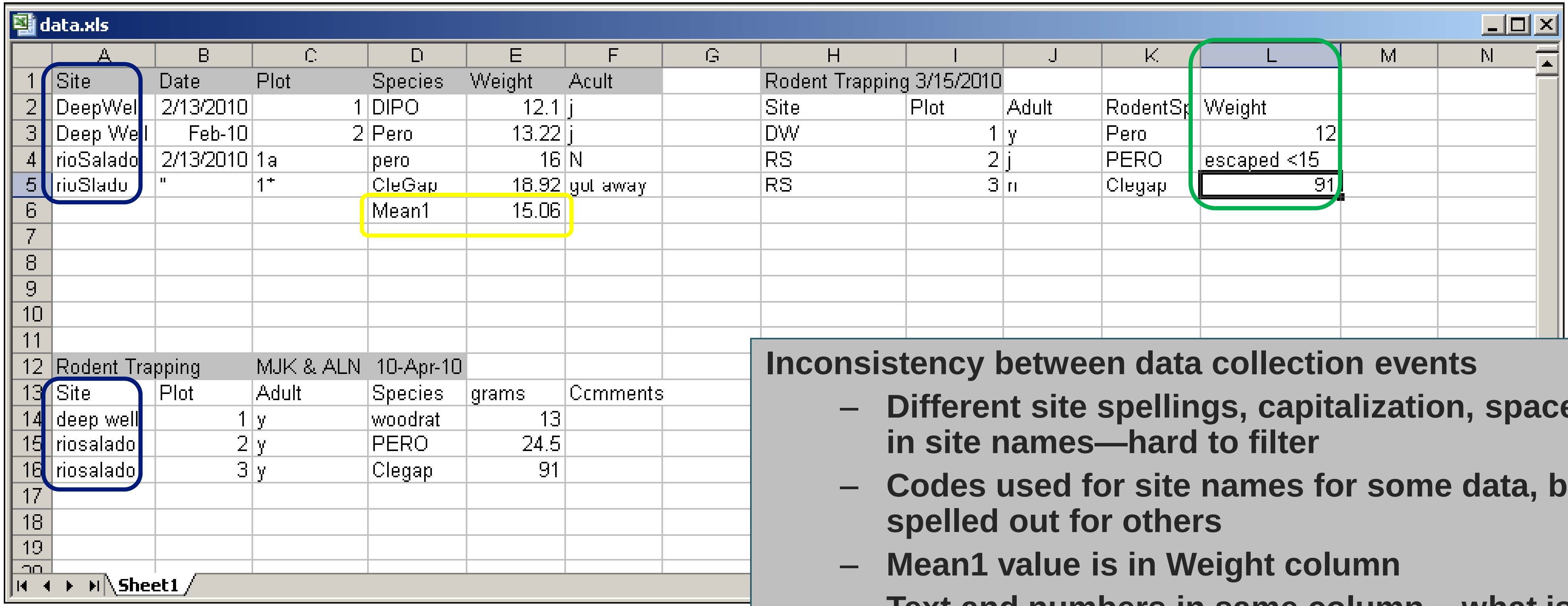
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Site	Date	Plot	Species	Weight	Acult		Rodent Trapping	3/15/2010					
2	DeepWell	2/13/2010	1	DIPO	12.1	j		Site	Plot	Adult	RodentSp	Weight		
3	Deep Wel	Feb-10	2	Pero	13.22	j		DW	1	y	Pero	12		
4	rioSalado	2/13/2010	1a	pero	16	N		RS	2	j	PERO	escaped <15		
5	riuSladu	"	1*	CleGap	18.92	gul away		RS	3	n	Clegap	91		
6				Mean1	15.06									
7														
8														
9														
10														
11														
12	Rodent Trapping		MJK & ALN	10-Apr-10										
13	Site	Plot	Adult	Species	grams	Ccmments								
14	deep well	1	y	woodrat	13									
15	riosalado	2	y	PERO	24.5									
16	riosalado	3	y	Clegap	91									
17														
18														
19														

Inconsistency between data collection events

- Location of Date information
- Inconsistent Date format
- Column names
- Order of columns

Example: unstructured data entry

From a small mammal trapping study

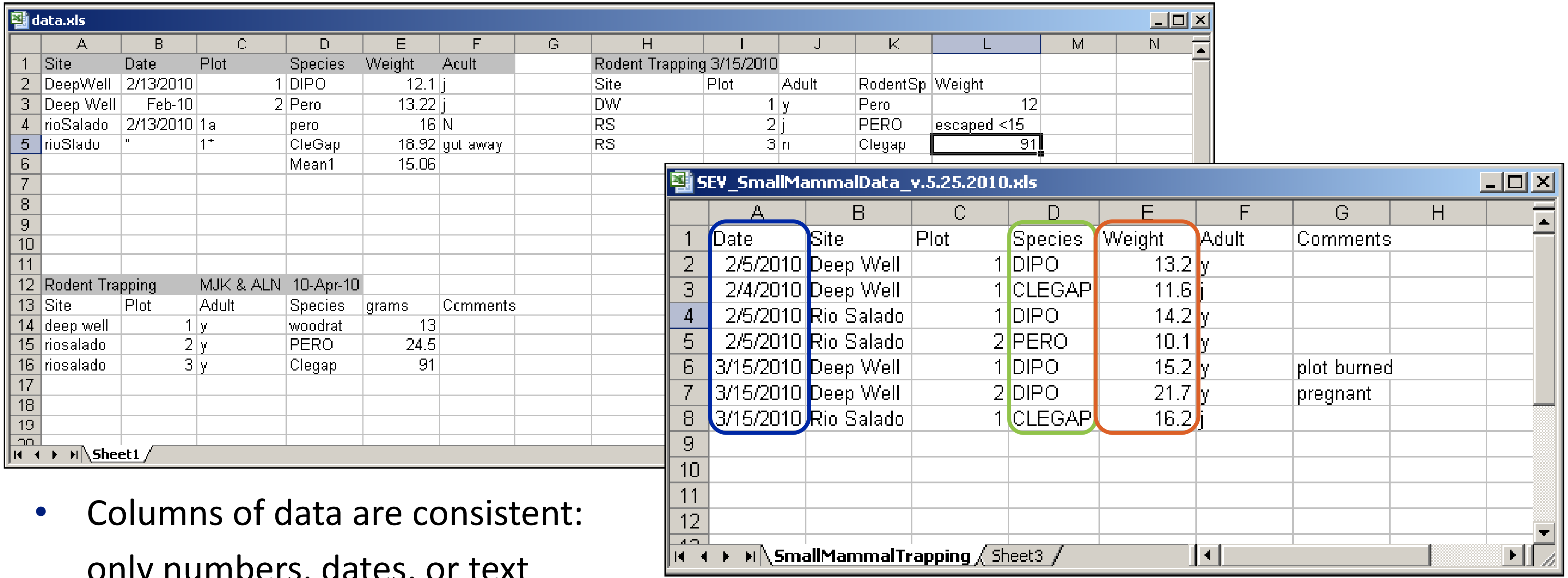


	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Site	Date	Plot	Species	Weight	Acult		Rodent Trapping 3/15/2010						
2	DeepWel	2/13/2010	1	DIPO	12.1	j		Site	Plot	Adult	RodentSp	Weight		
3	Deep Well	Feb-10	2	Pero	13.22	j		DW	1	y	Pero	12		
4	rioSalado	2/13/2010	1a	pero	16	N		RS	2	j	PERO	escaped <15		
5	riuSladu	"	1*	CleGap	18.92	gul away		RS	3	n	Clegap	91		
6				Mean1	15.06									
7														
8														
9														
10														
11														
12	Rodent Trapping		MJK & ALN	10-Apr-10										
13	Site	Plot	Adult	Species	grams	Ccmments								
14	deep well	1	y	woodrat	13									
15	riosalado	2	y	PERO	24.5									
16	riosalado	3	y	Clegap	91									
17														
18														
19														

Inconsistency between data collection events

- Different site spellings, capitalization, spaces in site names—hard to filter
- Codes used for site names for some data, but spelled out for others
- Mean1 value is in Weight column
- Text and numbers in same column – what is the mean of 12, “escaped < 15”, and 91?

The same data entry can be structured into one table



data.xls

	A	B	C	D	E	F
1	Site	Date	Plot	Species	Weight	Acult
2	DeepWell	2/13/2010	1	DIPO	12.1	j
3	Deep Well	Feb-10	2	Pero	13.22	j
4	rioSalado	2/13/2010	1a	pero	16	N
5	riuSladu	"	1*	CleGap	18.92	gul away
6				Mean1	15.06	

	H	I	J	K	L
1	Rodent Trapping 3/15/2010				
2	Site	Plot	Adult	RodentSp	Weight
3	DW	1	y	Pero	12
4	RS	2	j	PERO	escaped <15
5	RS	3	n	Clegap	91

	A	B	C	D	E	F
12	Rodent Trapping		MJK & ALN	10-Apr-10		
13	Site	Plot	Adult	Species	grams	Ccmments
14	deep well	1	y	woodrat	13	
15	riosalado	2	y	PERO	24.5	
16	riosalado	3	y	Clegap	91	

Sheet1

SEV_SmallMammalData_v.5.25.2010.xls

	A	B	C	D	E	F	G	H
1	Date	Site	Plot	Species	Weight	Adult	Comments	
2	2/5/2010	Deep Well	1	DIPO	13.2	y		
3	2/4/2010	Deep Well	1	CLEGAP	11.6	j		
4	2/5/2010	Rio Salado	1	DIPO	14.2	y		
5	2/5/2010	Rio Salado	2	PERO	10.1	y		
6	3/15/2010	Deep Well	1	DIPO	15.2	y	plot burned	
7	3/15/2010	Deep Well	2	DIPO	21.7	y	pregnant	
8	3/15/2010	Rio Salado	1	CLEGAP	16.2	j		

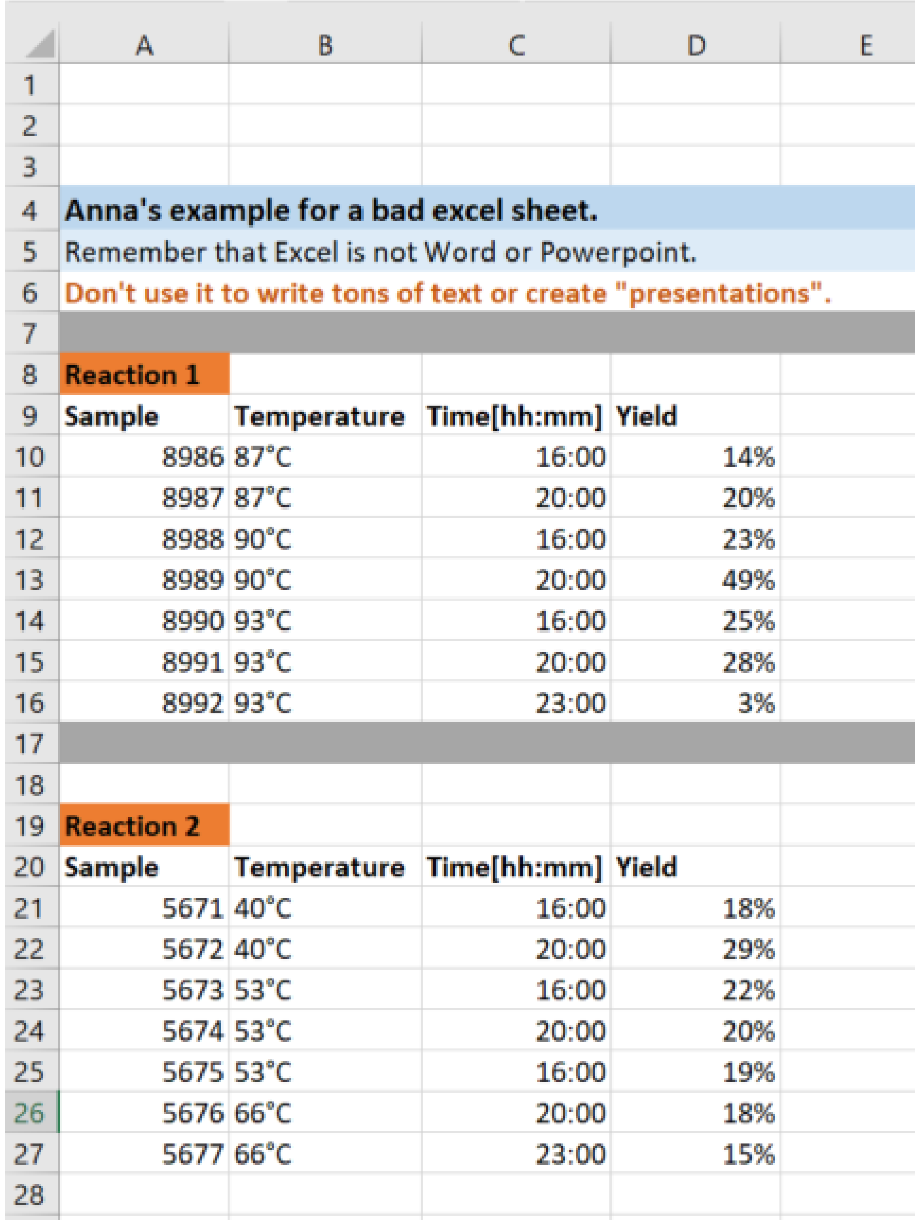
SmallMammalTrapping Sheet3

- Columns of data are consistent:

  only numbers, dates, or text
- Consistent Names, Codes, Formats (date) used in each column
- **Data are all in one table**, which is much easier for a statistical program to work with than multiple small tables which each require human intervention

Anna’s Excel-Tipps #1

- Always use the first line (A1, B1, C1, etc.) for the **column titles** and start entering data in cell A2.
- Do not create empty «interruptions lines»
- Do not start a new table on the same sheet.



	A	B	C	D	E
1					
2					
3					
4	**Anna's example for a bad excel sheet.**				
5	Remember that Excel is not Word or Powerpoint.				
6	Don't use it to write tons of text or create "presentations".				
7					
8	**Reaction 1**				
9	**Sample**	**Temperature**	**Time[hh:mm]**	**Yield**	
10	8986	87°C	16:00	14%	
11	8987	87°C	20:00	20%	
12	8988	90°C	16:00	23%	
13	8989	90°C	20:00	49%	
14	8990	93°C	16:00	25%	
15	8991	93°C	20:00	28%	
16	8992	93°C	23:00	3%	
17					
18					
19	**Reaction 2**				
20	**Sample**	**Temperature**	**Time[hh:mm]**	**Yield**	
21	5671	40°C	16:00	18%	
22	5672	40°C	20:00	29%	
23	5673	53°C	16:00	22%	
24	5674	53°C	20:00	20%	
25	5675	53°C	16:00	19%	
26	5676	66°C	20:00	18%	
27	5677	66°C	23:00	15%	
28					

Best practices for tables and spreadsheets

- Create descriptive column headers

- Careful with special characters!
  → Use **UTF-8 character encoding** when exporting or importing data if you use special characters!

- **Units**: Specify them in the column header or in a separate line under the header (some programs also have a dedicated line for the units).

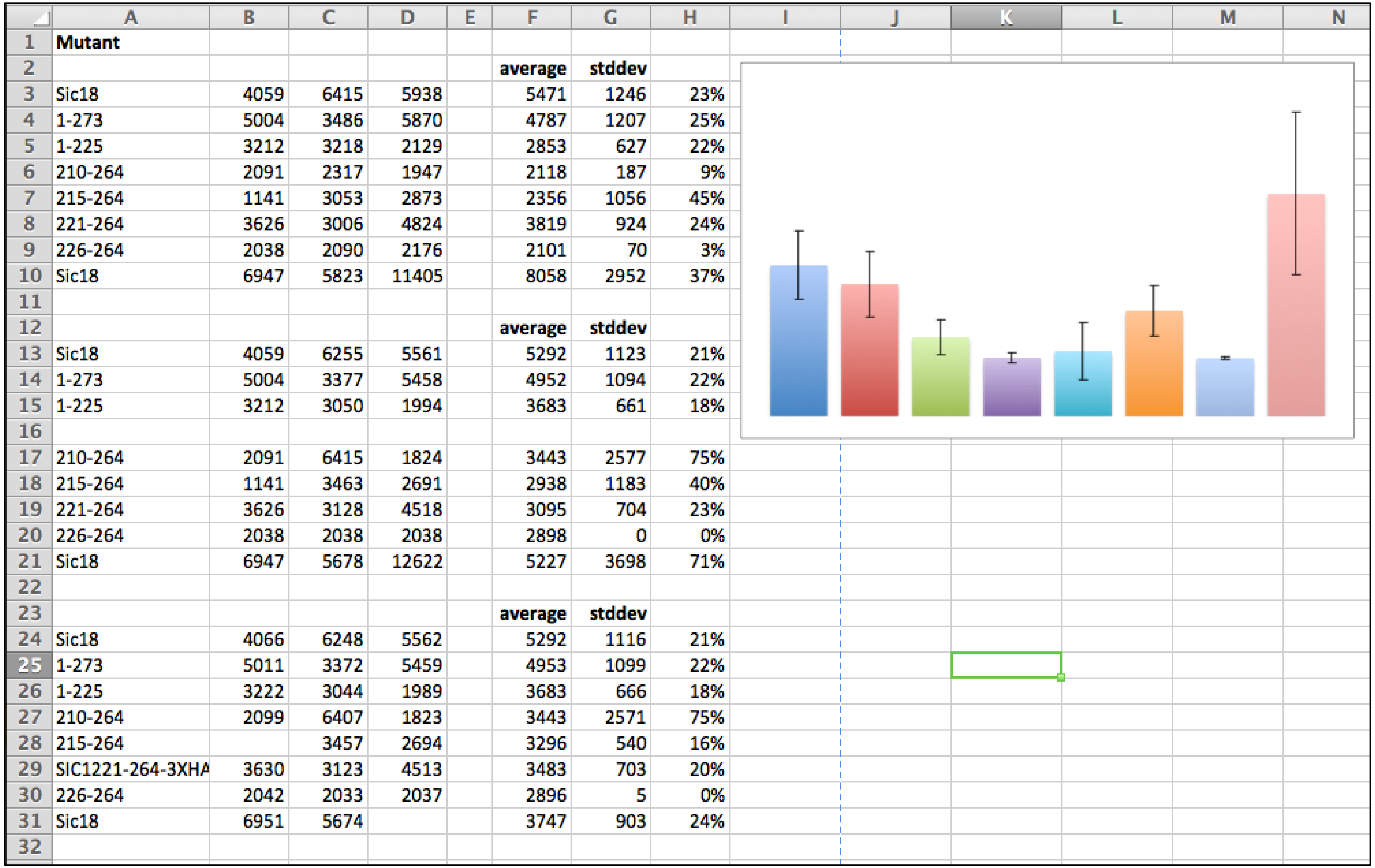
- Avoid empty spaces, many programs have problems to read them.
  → **Underlines** are the solution: «length m» → «length_m»

- Use uniform abbreviations and naming conventions throughout the spreadsheet

- Missing data: Leave field empty or create an abbreviation that indicates missing data. (depends on the software you use and how it handles empty fields)

Your turn:
Spot the *faux pas*!

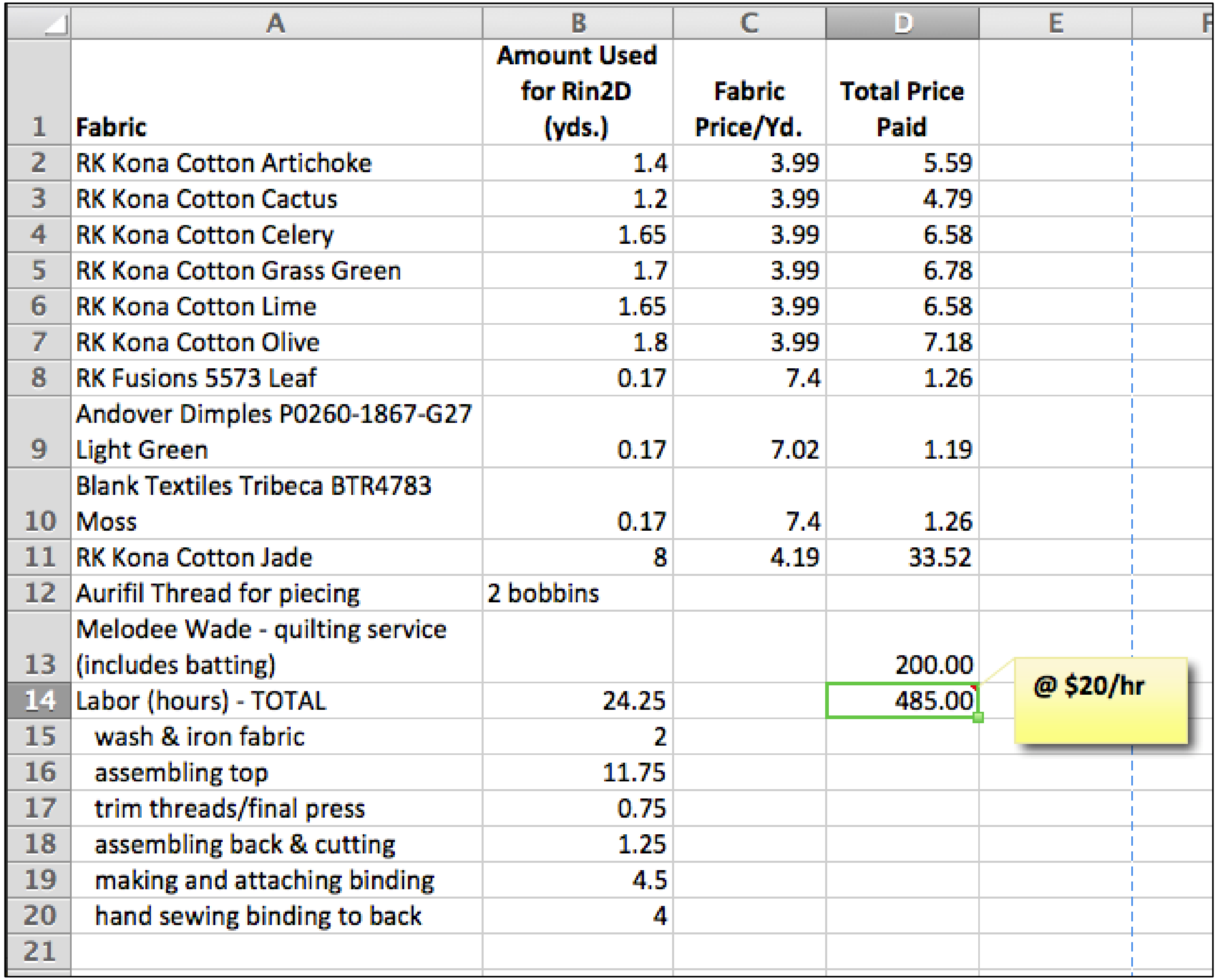
Exercise 3.1: Spot the six problems



	A	B	C	D	E	F	G	H
1	**Mutant**							
2						**average**	**stddev**	
3	Sic18	4059	6415	5938		5471	1246	23%
4	1-273	5004	3486	5870		4787	1207	25%
5	1-225	3212	3218	2129		2853	627	22%
6	210-264	2091	2317	1947		2118	187	9%
7	215-264	1141	3053	2873		2356	1056	45%
8	221-264	3626	3006	4824		3819	924	24%
9	226-264	2038	2090	2176		2101	70	3%
10	Sic18	6947	5823	11405		8058	2952	37%
11								
12						**average**	**stddev**	
13	Sic18	4059	6255	5561		5292	1123	21%
14	1-273	5004	3377	5458		4952	1094	22%
15	1-225	3212	3050	1994		3683	661	18%
16								
17	210-264	2091	6415	1824		3443	2577	75%
18	215-264	1141	3463	2691		2938	1183	40%
19	221-264	3626	3128	4518		3095	704	23%
20	226-264	2038	2038	2038		2898	0	0%
21	Sic18	6947	5678	12622		5227	3698	71%
22								
23						**average**	**stddev**	
24	Sic18	4066	6248	5562		5292	1116	21%
25	1-273	5011	3372	5459		4953	1099	22%
26	1-225	3222	3044	1989		3683	666	18%
27	210-264	2099	6407	1823		3443	2571	75%
28	215-264		3457	2694		3296	540	16%
29	SIC1221-264-3XHA	3630	3123	4513		3483	703	20%
30	226-264	2042	2033	2037		2896	5	0%
31	Sic18	6951	5674			3747	903	24%
32								

Exercise 3.1: Six problems

Exercise 3.2: Spot the two problems



	A	B	C	D	E	F
1	Fabric	Amount Used for Rin2D (yds.)	Fabric Price/Yd.	Total Price Paid		
2	RK Kona Cotton Artichoke	1.4	3.99	5.59		
3	RK Kona Cotton Cactus	1.2	3.99	4.79		
4	RK Kona Cotton Celery	1.65	3.99	6.58		
5	RK Kona Cotton Grass Green	1.7	3.99	6.78		
6	RK Kona Cotton Lime	1.65	3.99	6.58		
7	RK Kona Cotton Olive	1.8	3.99	7.18		
8	RK Fusions 5573 Leaf	0.17	7.4	1.26		
9	Andover Dimples P0260-1867-G27 Light Green	0.17	7.02	1.19		
10	Blank Textiles Tribeca BTR4783 Moss	0.17	7.4	1.26		
11	RK Kona Cotton Jade	8	4.19	33.52		
12	Aurifil Thread for piecing	2 bobbins				
13	Melodee Wade - quilting service (includes batting)			200.00		
14	Labor (hours) - TOTAL	24.25		485.00		
15	wash & iron fabric	2				
16	assembling top	11.75				
17	trim threads/final press	0.75				
18	assembling back & cutting	1.25				
19	making and attaching binding	4.5				
20	hand sewing binding to back	4				
21						

@ $20/hr

Exercise 3.2: Two problems

Summary: Structure your data!

- **Are your data as structured as possible?**
- **Integrate as much data as possible into tables / spreadsheets**
- **Combine tables / spreadsheets whenever possible**

But I work with images / audio / video data.
What can I do?

That’s ok. Research can also rely on unstructured data. Lesson 6 «Data Documentation and Metadata» will be very important for the management of your data.

Lesson 3: Data Entry and Manipulation

✓ How to structure your data: Best practices

→Quality of research data

- Data entry tools
- Databases
- Data Analysis

Types of «bad research data»

- Inconsistent / unreliable data
- Invalid / Inaccurate data

- Incomplete data

- Nonintegrated data

Research data quality characteristics

Reliability ~ Consistency ~ Reproducibility

The extent to which **the results can be reproduced** when the research is repeated under the same conditions.

Assessed by checking the consistency of results across time, across different observers, and across parts of the test itself.

A reliable measurement is not always valid: the results might be reproducible, but they're not necessarily correct.

Validity ~ Accuracy

The extent to which the results **really measure what they are supposed to measure.**

Assessed by checking how well the results correspond to established theories and other measures of the same concept.

A valid measurement is generally reliable: if a test produces accurate results, they should be reproducible.

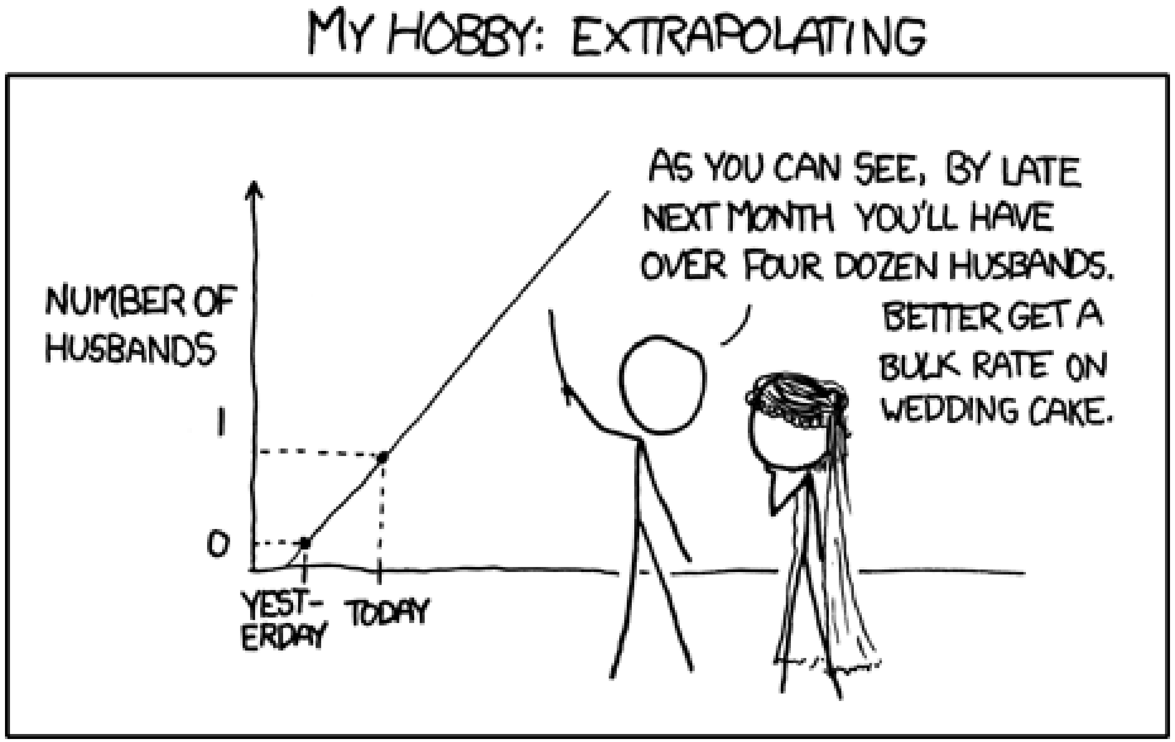
Scribbr. Reliability vs Validity in Research | Differences, Types and Examples.
https://www.scribbr.com/methodology/reliability-vs-validity/
accessed: Aug 26th 2020

Research data quality characteristics

Completeness

There are two kinds of people:
1) Those who cannot extrapolate from incomplete data

https://www.teezily.com/


MY HOBBY: EXTRAPOLATING
NUMBER OF HUSBANDS
1
0
YEST-ERDAY
TODAY
AS YOU CAN SEE, BY LATE NEXT MONTH YOU'LL HAVE OVER FOUR DOZEN HUSBANDS. BETTER GET A BULK RATE ON WEDDING CAKE.

https://xkcd.com/605/

While extrapolation is often useful, it might not always get you accurate results…
So **make sure your datasets are as complete as possible!**

Research data quality characteristics

Data integration = process of combining data from different sources into a single unified view

Typically required for

- Business intelligence
- Big Data analyses

  What is Data Integration? | Talend https://www.talend.com/resources/what-is-data-integration/

- when reusing research data

Lesson 3: Data Entry and Manipulation

✓ How to structure your data: Best practices

✓ Quality of research data

→ Data entry tools

- Databases
- Data Analysis

Data entry tools

For Spreadsheets

MS Excel

Apple Numbers

Google Sheets

OpenOffice Calc

LibreOffice Calc

Zoho Sheets

Data entry tools

For Surveys

LimeSurvey

Surveymonkey

Scientific online survey tool
Campus licence available for all UZH members
https://www.uzh.ch/zi/cl/umfragen/index.php/admin/authentication/sa/login

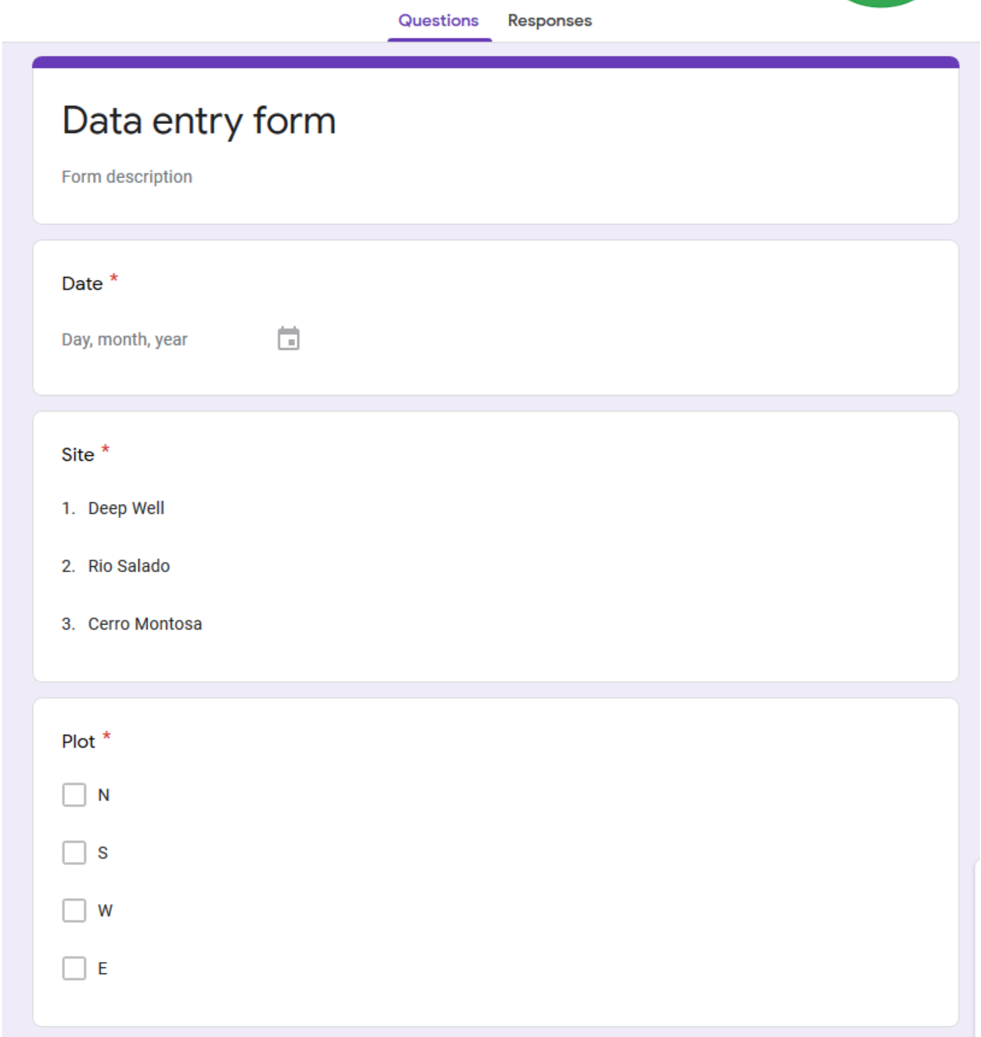
Google Forms

Google forms: Not only for surveys!

- Enter data through a form
- Can be directly fed into a Google spreadsheet

- **Pros**:
  - Predefined answer possibilities («controlled vocabulary») → Data validation
  - Easier to receive a well-structured spreadsheet

- **Cons**:
  - Doesn't work well with validation of numerical values (e.g. numbers only in a certain range)



Data entry form

Form description

Date *

Day, month, year

Site *

1. Deep Well
2. Rio Salado
3. Cerro Montosa

Plot *

□ N

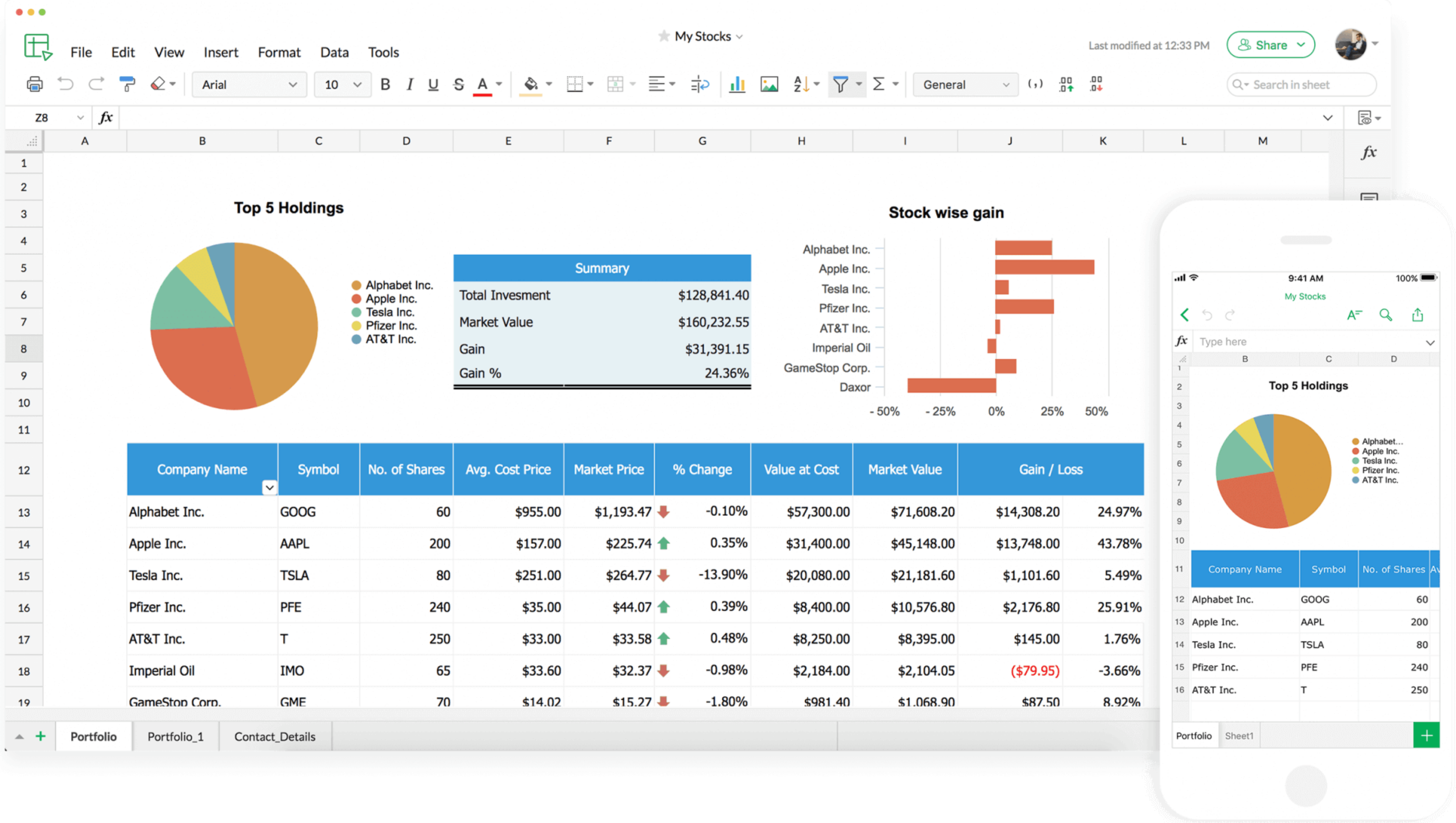
□ S

□ W

□ E

Zoho Sheets

- Similar to Google sheets, but much more functionality
- Data validation
- Data entry through forms
- Analysis tools,
  e.g. Pivot tables



Summary	
Total Invesment	$128,841.40
Market Value	$160,232.55
Gain	$31,391.15
Gain %	24.36%

Company Name	Symbol	No. of Shares	Avg. Cost Price	Market Price	% Change	Value at Cost	Market Value	Gain / Loss	
Alphabet Inc.	GOOG	60	$955.00	$1,193.47	-0.10%	$57,300.00	$71,608.20	$14,308.20	24.97%
Apple Inc.	AAPL	200	$157.00	$225.74	0.35%	$31,400.00	$45,148.00	$13,748.00	43.78%
Tesla Inc.	TSLA	80	$251.00	$264.77	-13.90%	$20,080.00	$21,181.60	$1,101.60	5.49%
Pfizer Inc.	PFE	240	$35.00	$44.07	0.39%	$8,400.00	$10,576.80	$2,176.80	25.91%
AT&T Inc.	T	250	$33.00	$33.58	0.48%	$8,250.00	$8,395.00	$145.00	1.76%
Imperial Oil	IMO	65	$33.60	$32.37	-0.98%	$2,184.00	$2,104.05	($79.95)	-3.66%
GameStop Corp.	GME	70	$14.02	$15.27	-1.80%	$981.40	$1,068.90	$87.50	8.92%

Anna’s Excel-Tipps #2

Demo data validation:

How to predefine answer possibilities in Excel

– Data → Data validation

Lesson 3: Data Entry and Manipulation

✓ How to structure your data: Best practices

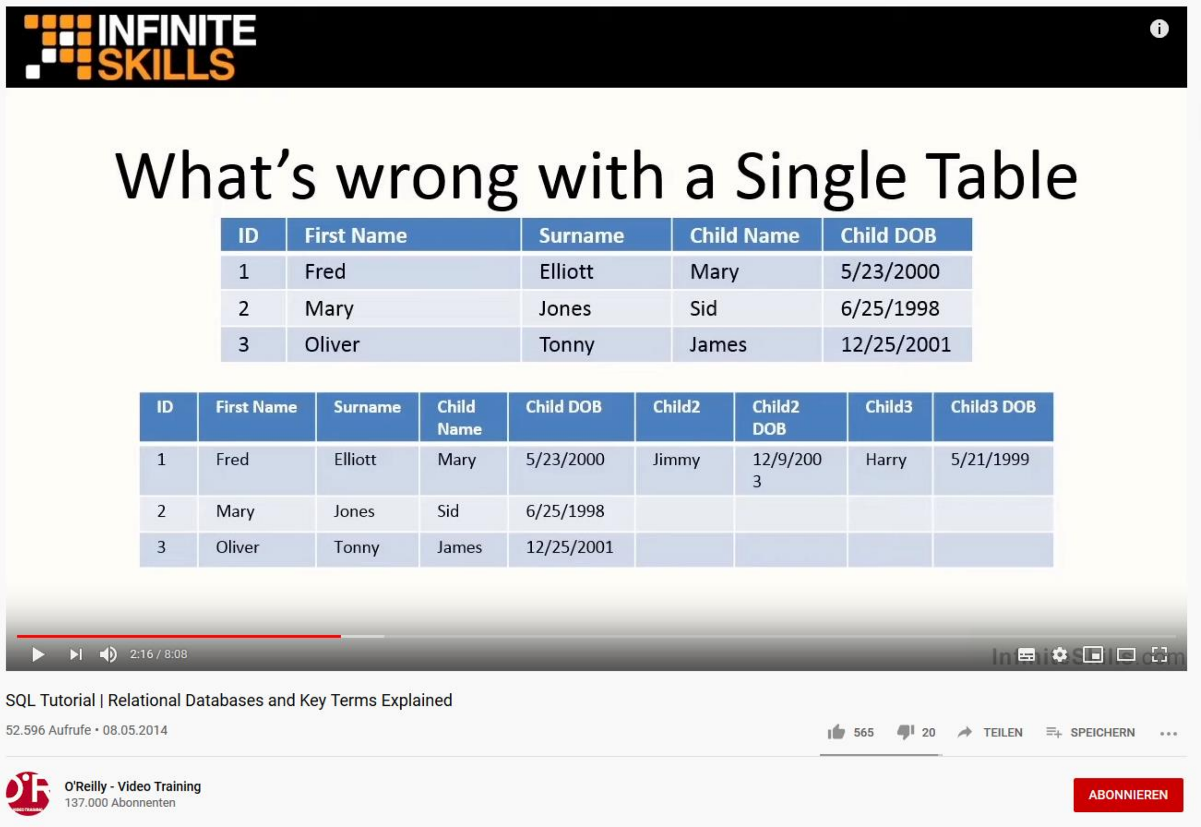
✓ Quality of research data

✓ Data entry tools

→ Databases

- Data Analysis

What’s wrong with a Single Table?



What’s wrong with a Single Table

ID	First Name	Surname	Child Name	Child DOB
1	Fred	Elliott	Mary	5/23/2000
2	Mary	Jones	Sid	6/25/1998
3	Oliver	Tonny	James	12/25/2001

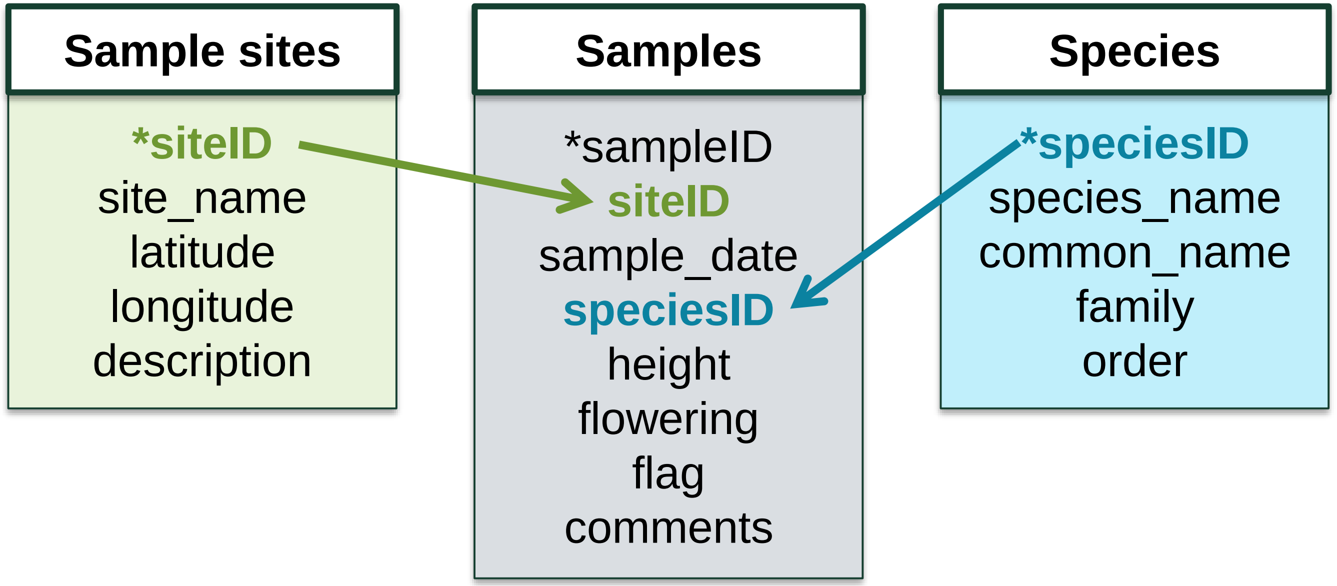
ID	First Name	Surname	Child Name	Child DOB	Child2	Child2 DOB	Child3	Child3 DOB
1	Fred	Elliott	Mary	5/23/2000	Jimmy	12/9/200 3	Harry	5/21/1999
2	Mary	Jones	Sid	6/25/1998				
3	Oliver	Tonny	James	12/25/2001				

SQL Tutorial | Relational Databases and Key Terms Explained

52.596 Aufrufe • 08.05.2014

https://youtu.be/h8IWmmxIyS0?t=83

What is a relational database?



Sample sites

*siteID
site_name
latitude
longitude
description

Samples

*sampleID
siteID
sample_date
speciesID
height
flowering
flag
comments

Species

*speciesID
species_name
common_name
family
order

- Contains more than one table
- Relationships between the tables
- Parent tables and child tables
- Are searched with a declarative programming language: **SQL = structured query language**

Spreadsheets vs. databases

Spreadsheets

- Flexible about cell content type—cells in same column can contain numbers or text
- Cells can contain calculations (functions)
- Limited number of rows
- usually not editable by multiple users at the same time
- Allow for extensive analysis

Databases

- Pre-set the type of data contained in a certain field
- Suitable for very large amounts of raw data
- Improved data integrity and consistency
- multiple users can work on it in parallel
- All calculations and operations are done after data retrieval

Want to give databases a try?

- MySQL (open-source, aquired in 2010 by Oracle)
- **MariaDB** (fork of MySQL)
  - community developed
  - Intended to remain free and open-source under GNU GPL

  - Tutorials to get started: https://mariadb.com/get-started-with-mariadb/
  - Geographic & Geometric Features in MariaDB: https://mariadb.com/kb/en/geographic-geometric-features/

Lesson 3: Data Entry and Manipulation

✓ How to structure your data: Best practices

✓ Quality of research data

✓ Data entry tools

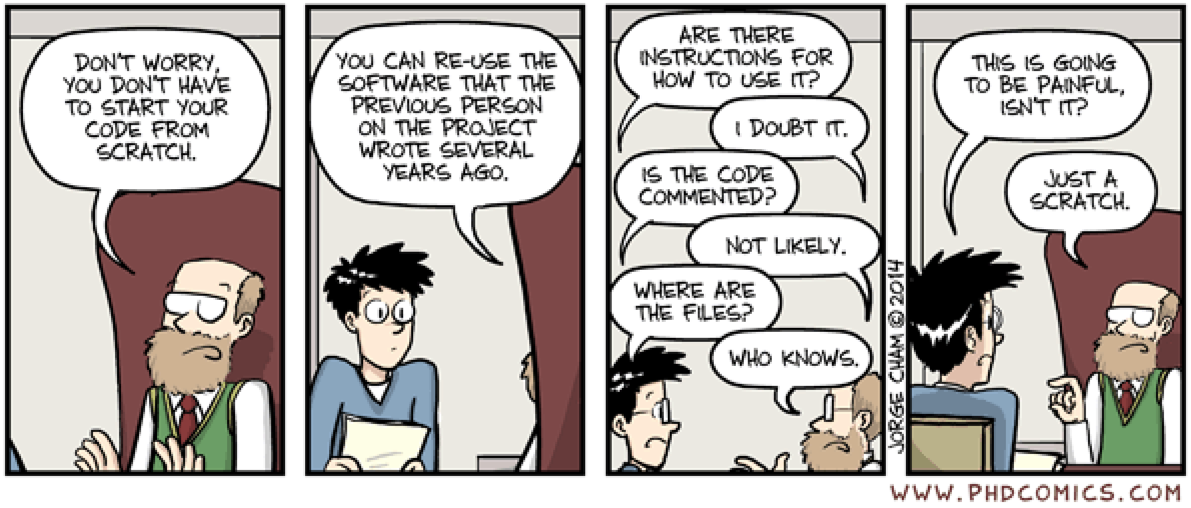
✓ Databases

→Data Analysis

Think of reproducibility when analysing data!

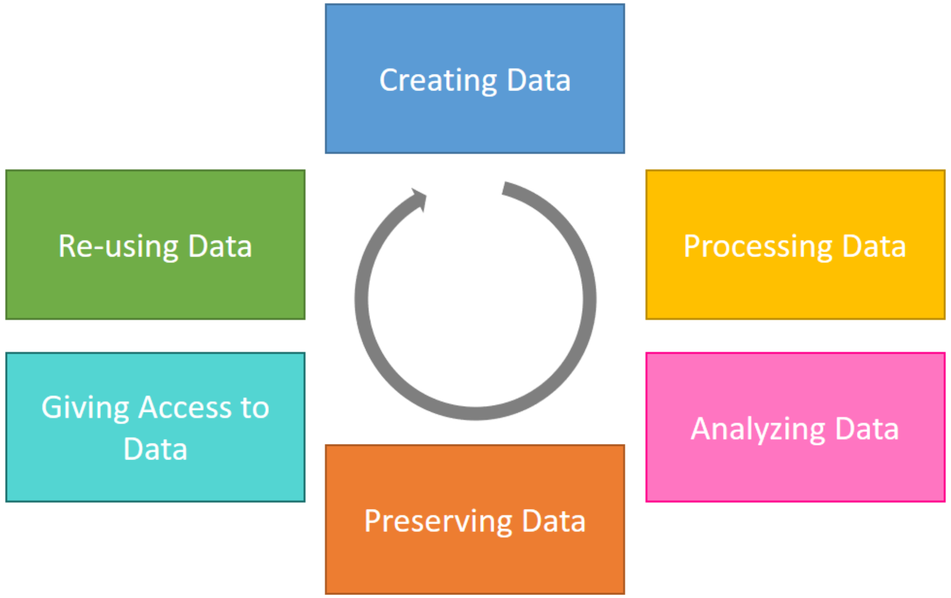
- **Document** your data analysis process
- «Metadata»: data about data
  - **Process metadata**: data documenting the process used to create, manipulate, and analyze data

→ Lesson 6: Data documentation & Metadata


DON'T WORRY, YOU DON'T HAVE TO START YOUR CODE FROM SCRATCH.
YOU CAN RE-USE THE SOFTWARE THAT THE PREVIOUS PERSON ON THE PROJECT WROTE SEVERAL YEARS AGO.
ARE THERE INSTRUCTIONS FOR HOW TO USE IT?
I DOUBT IT.
IS THE CODE COMMENTED?
NOT LIKELY.
WHERE ARE THE FILES?
WHO KNOWS.
JORGE CHAM © 2014
THIS IS GOING TO BE PAINFUL, ISN'T IT?
JUST A SCRATCH.
WWW.PHDCOMICS.COM

Data provenance

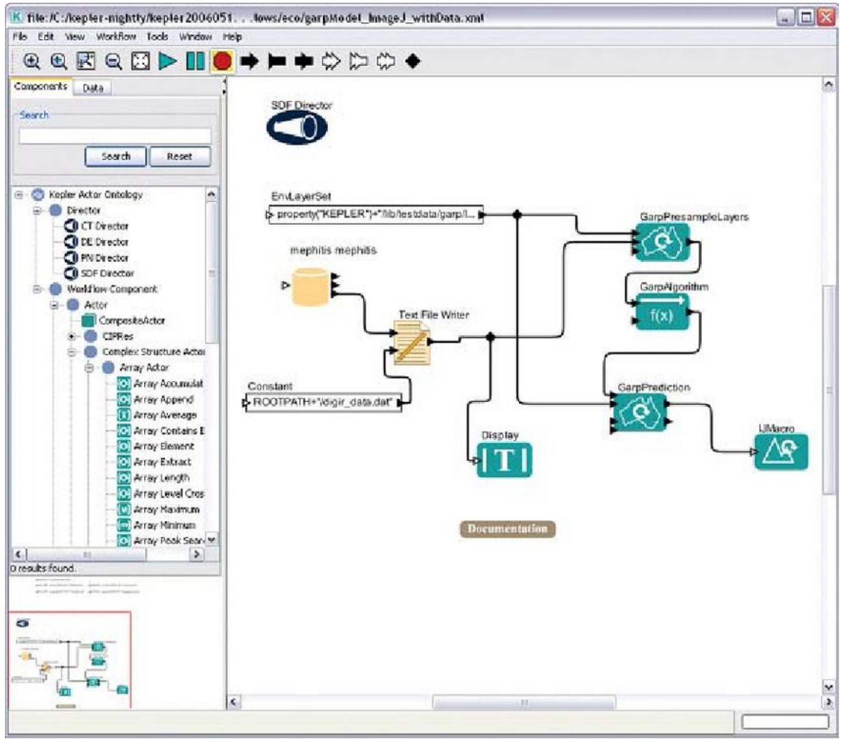
- Description of the origins of data
- Ability to follow data throughout the entire **life cycle**
  - Replication / reproducibility
  - Detection of **potential defects**, logical or statistical errors, **limitations**
  - Evaluation of hypotheses
- Especially important for making data **reusable**



Tools for documenting scientific workflows

kepler-project.org/

- Open-source, free, cross-platform
- Drag-and-drop interface for workflow construction
- Possible applications
  - Theoretical models or observational analyses
  - Hierarchical modeling
  - Can have nested workflows
  - Can access data from web-based sources (e.g. databases)



Summary of Lesson 3

Create **structured** data whenever possible

Make sure your data is **consistent**, **reproducible**, **accurate** and **complete**.

When using data from different sources: Make sure your data is well **integrated**.

Choose a data entry method that allows for the **validation** of data as it is entered.

Consider investing time in learning how to use a **relational database** if datasets are large or complex.

Remember to **document** your data analysis and manipulation to ensure **reusability** and **reproducibility**.