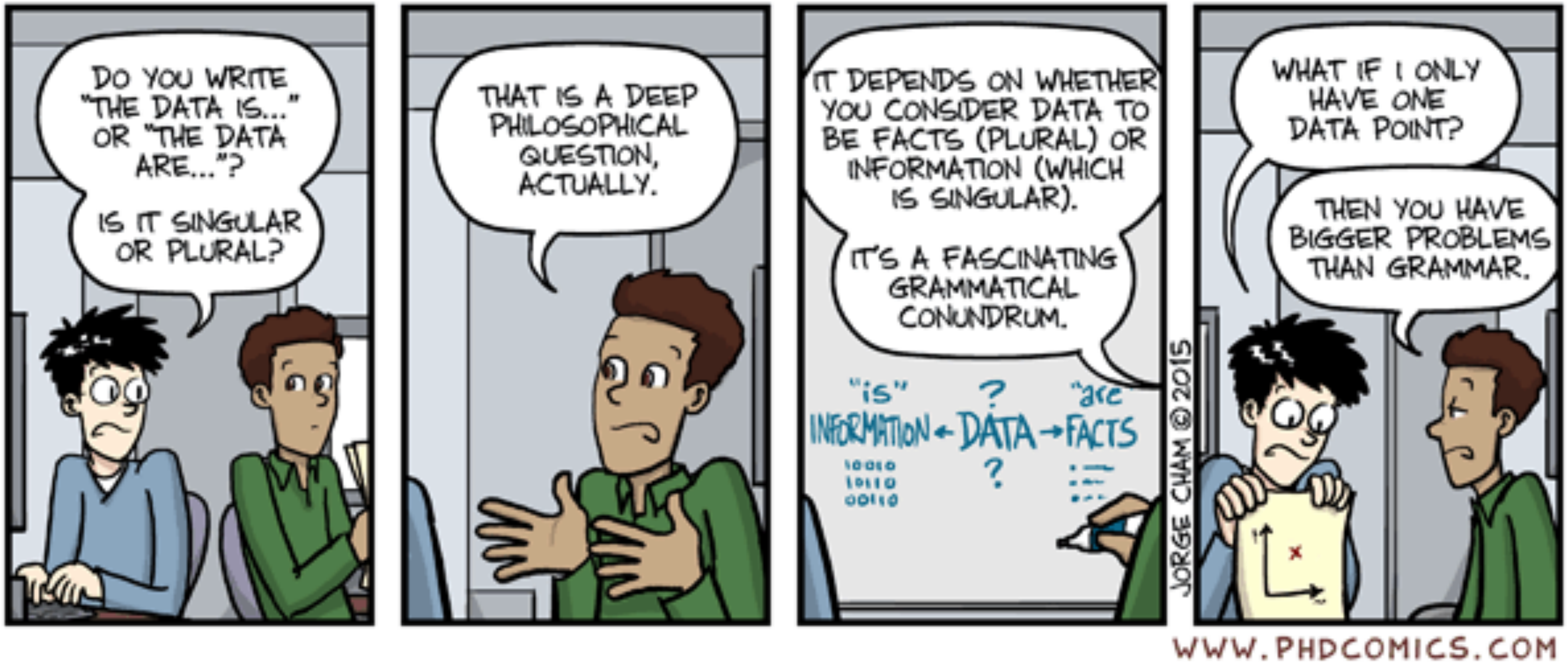
FHL 470
Spreadsheets and data management


DO YOU WRITE "THE DATA IS..." OR "THE DATA ARE..."?
IS IT SINGULAR OR PLURAL?
THAT IS A DEEP PHILOSOPHICAL QUESTION, ACTUALLY.
IT DEPENDS ON WHETHER YOU CONSIDER DATA TO BE FACTS (PLURAL) OR INFORMATION (WHICH IS SINGULAR).
IT'S A FASCINATING GRAMMATICAL CONUNDRUM.
"is"
?
"are"
INFORMATION ← DATA → FACTS
?
10010
10110
00110
WHAT IF I ONLY HAVE ONE DATA POINT?
THEN YOU HAVE BIGGER PROBLEMS THAN GRAMMAR.
JORGE CHAM © 2015
WWW.PHDCOMICS.COM

Robin Elahi, Hilary Hayford
Friday Harbor Labs
University of Washington

Taylor & Francis
Taylor & Francis Group

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: http://www.tandfonline.com/loi/utas20

Data organization in spreadsheets

Karl W. Broman & Kara H. Woo

To cite this article: Karl W. Broman & Kara H. Woo (2017): Data organization in spreadsheets, The American Statistician, DOI: 10.1080/00031305.2017.1375989

To link to this article: http://dx.doi.org/10.1080/00031305.2017.1375989

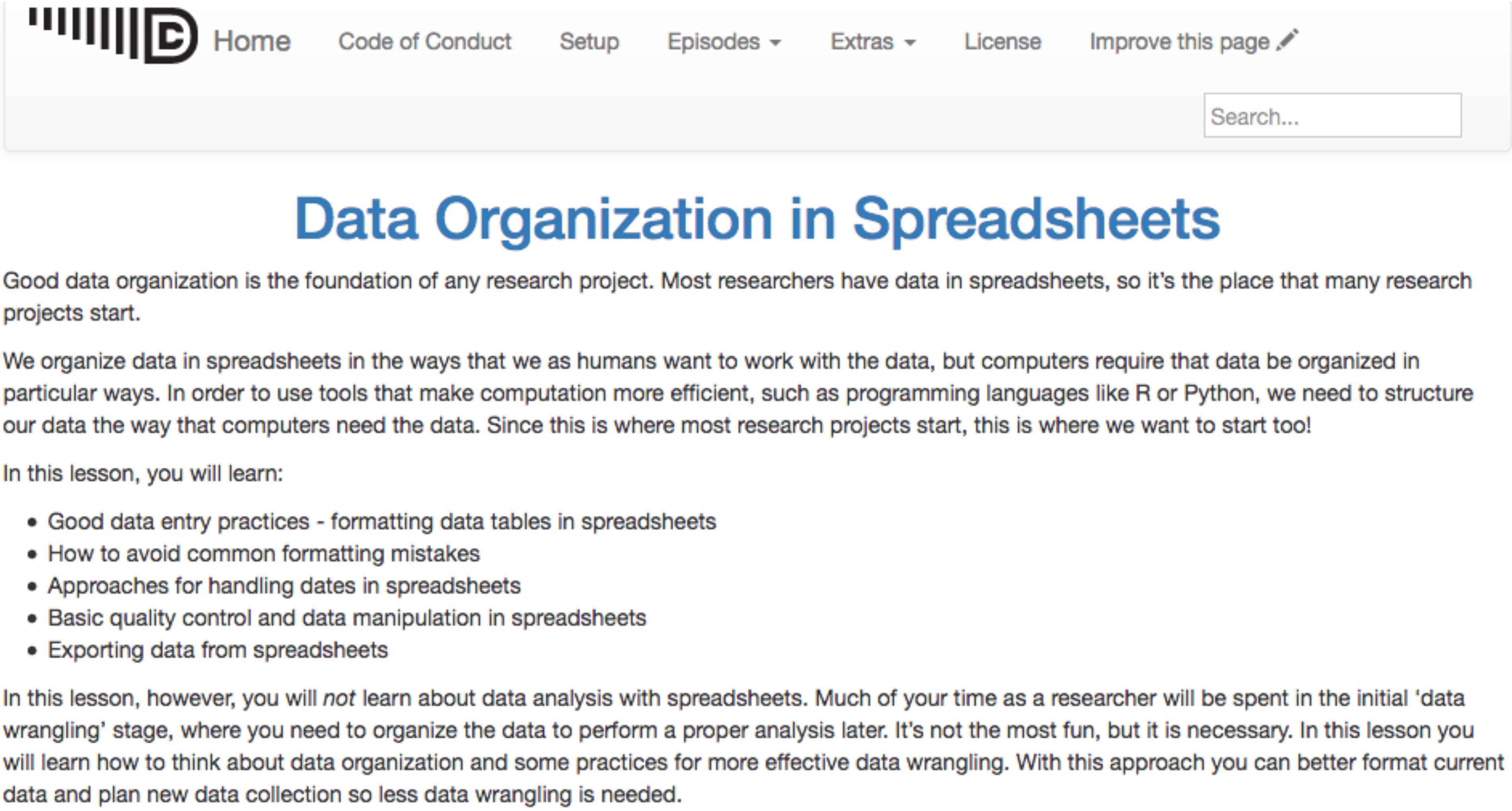
The full slide deck may be downloaded from:

http://www.dataone.org/education-modules

Suggested citation:

DataONE Education Module: Data Entry and Manipulation. DataONE. Retrieved Nov12, 2012. From http://www.dataone.org/sites/all/documents/L04_DataEntryManipulation.pptx

Copyright license information:



Data Organization in Spreadsheets

Good data organization is the foundation of any research project. Most researchers have data in spreadsheets, so it's the place that many research projects start.

We organize data in spreadsheets in the ways that we as humans want to work with the data, but computers require that data be organized in particular ways. In order to use tools that make computation more efficient, such as programming languages like R or Python, we need to structure our data the way that computers need the data. Since this is where most research projects start, this is where we want to start too!

In this lesson, you will learn:

- Good data entry practices - formatting data tables in spreadsheets
- How to avoid common formatting mistakes
- Approaches for handling dates in spreadsheets
- Basic quality control and data manipulation in spreadsheets
- Exporting data from spreadsheets

In this lesson, however, you will *not* learn about data analysis with spreadsheets. Much of your time as a researcher will be spent in the initial 'data wrangling' stage, where you need to organize the data to perform a proper analysis later. It's not the most fun, but it is necessary. In this lesson you will learn how to think about data organization and some practices for more effective data wrangling. With this approach you can better format current data and plan new data collection so less data wrangling is needed.

http://www.datacarpentry.org/spreadsheet-ecology-lesson/

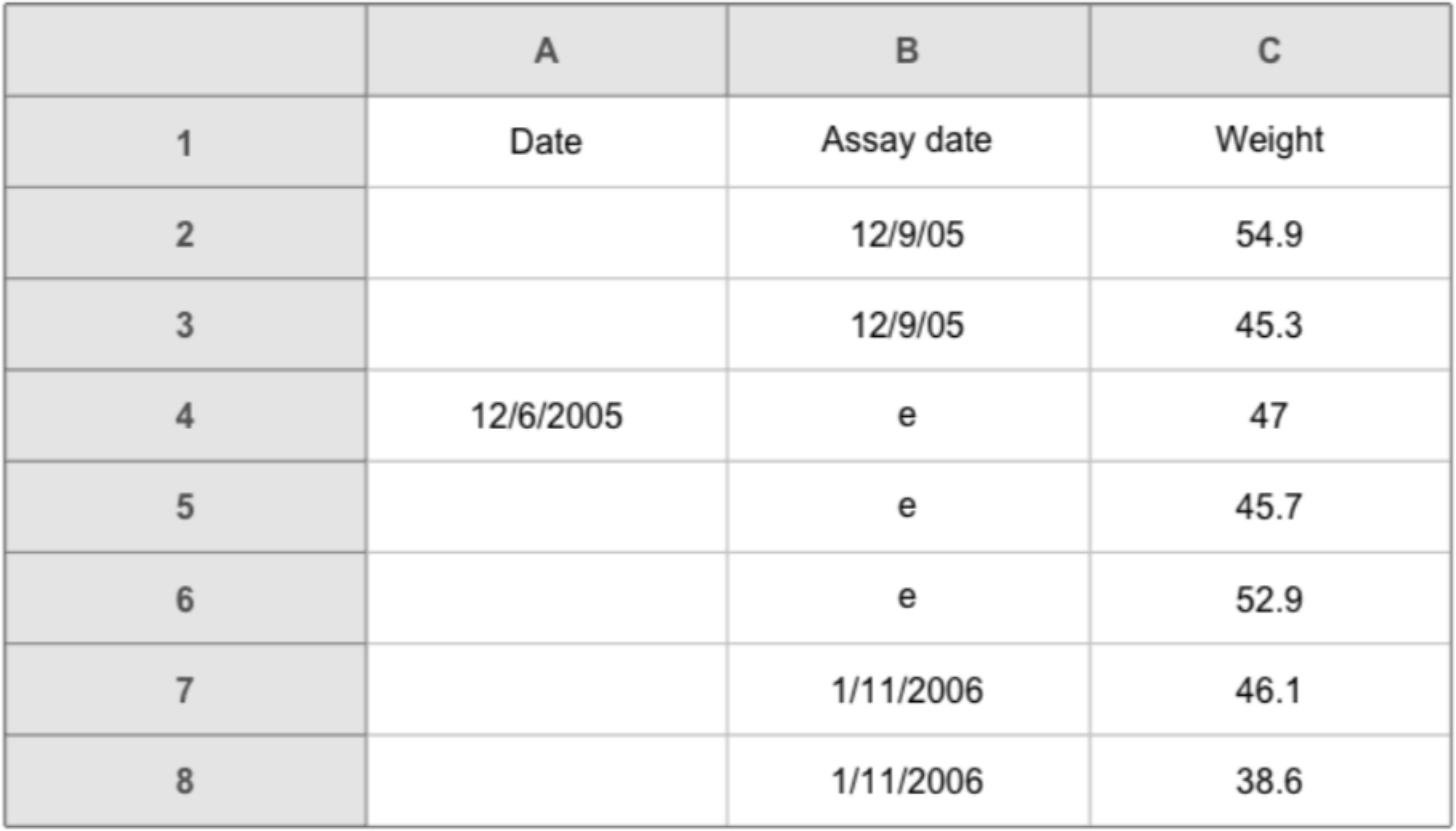
“Spreadsheets, for all of their mundane rectangularness, have been the subject of controversy for decades”

Broman and Woo 2018

1. Be consistent
2. Choose good names for things
3. Write dates as YYYY-MM-DD
4. No empty cells
5. Just one thing in a cell
6. Make it a rectangle
7. Create a data dictionary
8. No calculations in raw data files
9. Don’t use font color or highlighting as data
10. Make backups
11. Use data validation to avoid data entry errors
12. Save the data as plain text (.txt or .csv)

Broman and Woo 2018

Be consistent

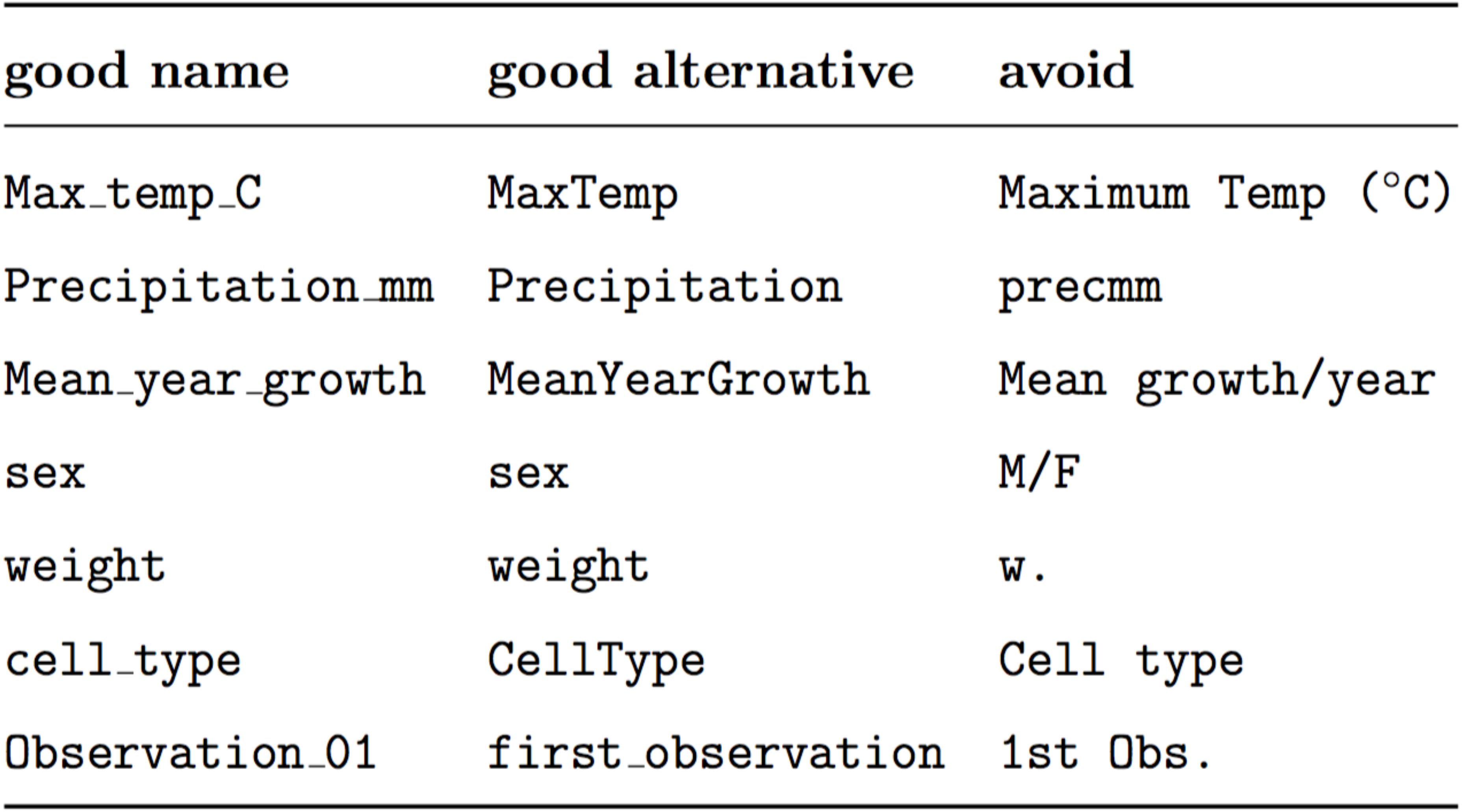


	A	B	C
1	Date	Assay date	Weight
2		12/9/05	54.9
3		12/9/05	45.3
4	12/6/2005	e	47
5		e	45.7
6		e	52.9
7		1/11/2006	46.1
8		1/11/2006	38.6

Choose good names for things

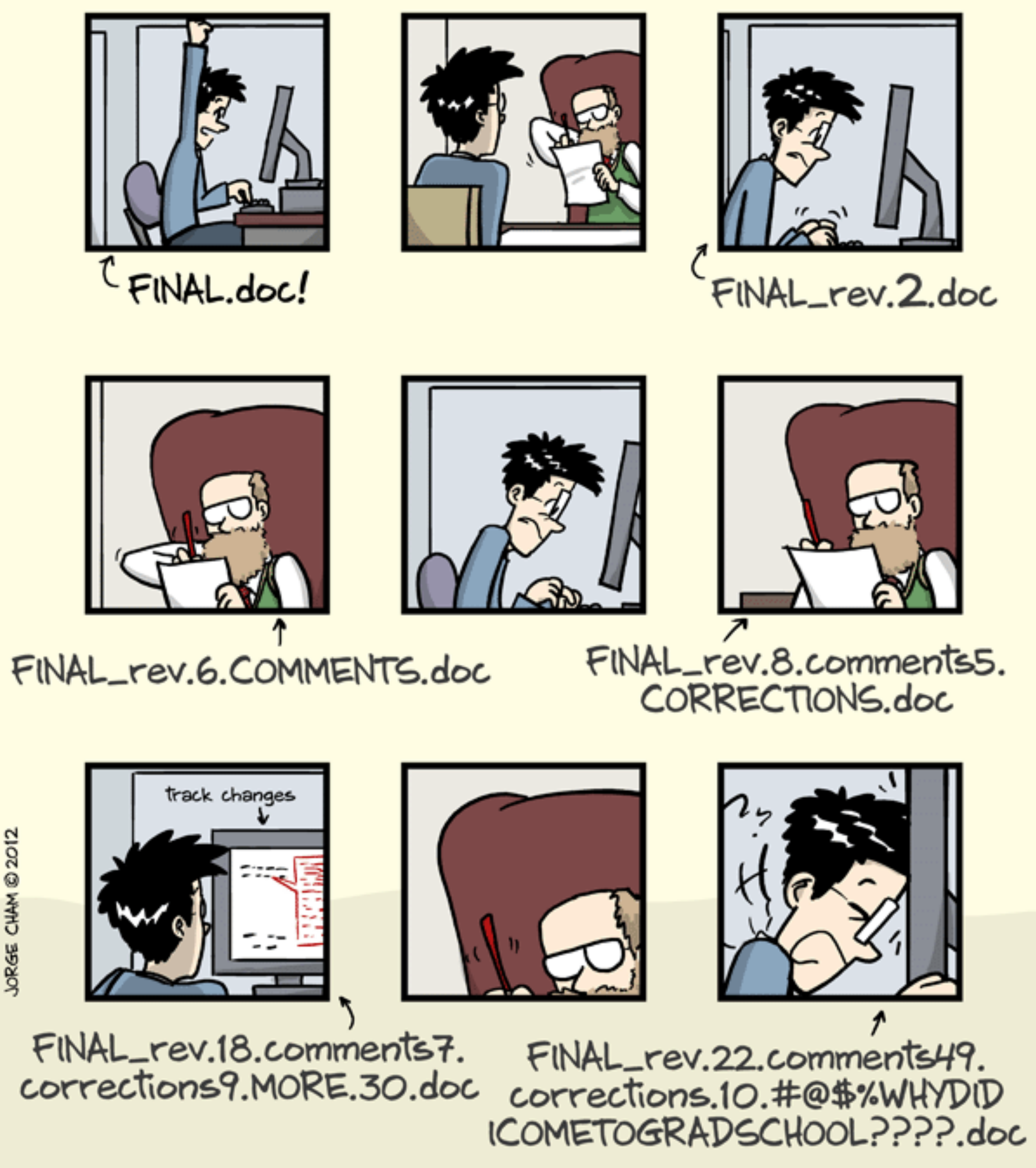
- Create descriptive column names without spaces or special characters
  - Soil T30 → Soil_Temp_30cm
  - Species-Code → Species_Code
  - Avoid using -,+,*,^, /, $, @, &, %, etc. in column names)

Choose good names for things



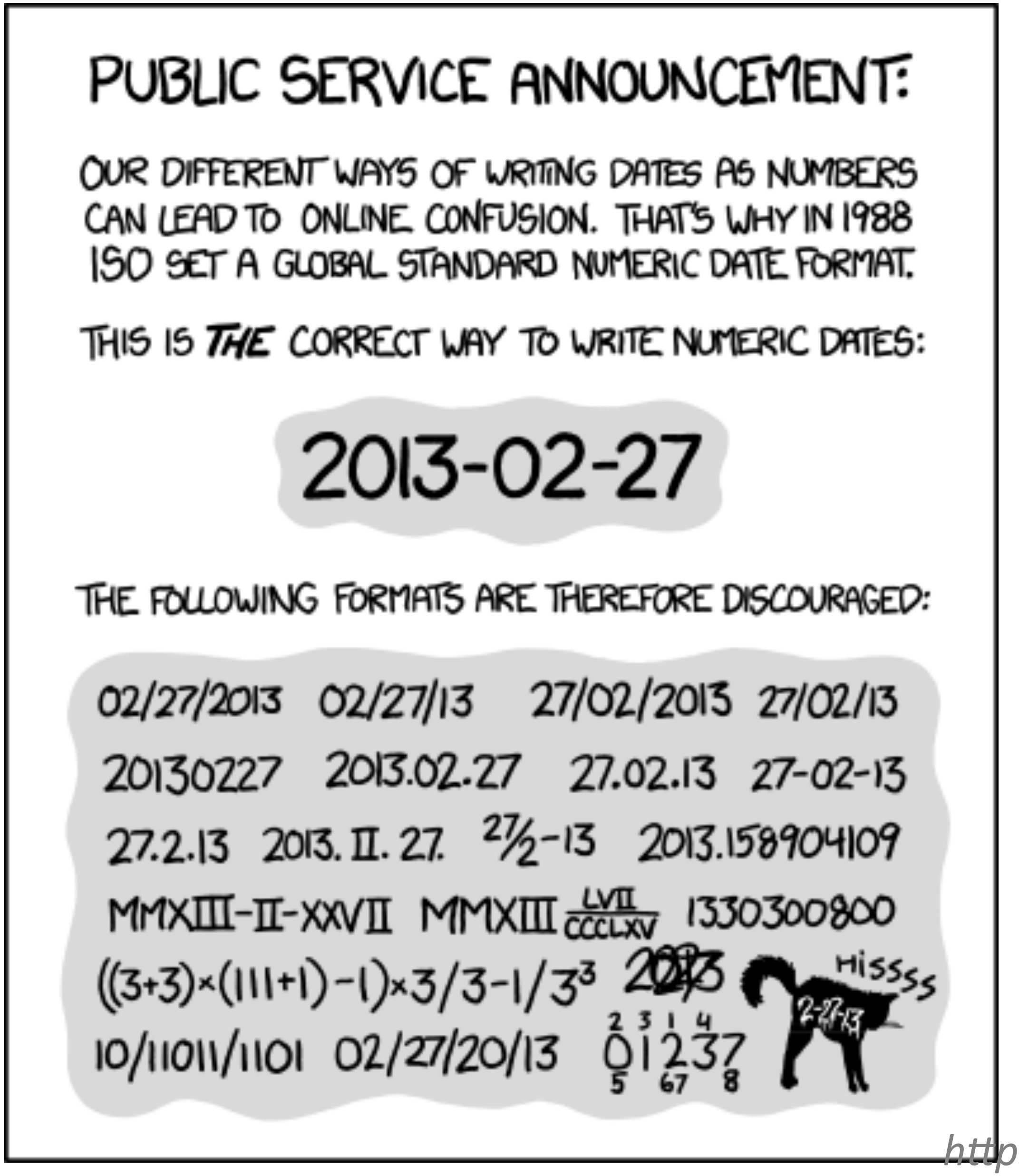
good name	good alternative	avoid
Max_temp_C	MaxTemp	Maximum Temp (°C)
Precipitation_mm	Precipitation	precmm
Mean_year_growth	MeanYearGrowth	Mean growth/year
sex	sex	M/F
weight	weight	w.
cell_type	CellType	Cell type
Observation_01	first_observation	1st Obs.

Broman and Woo 2018

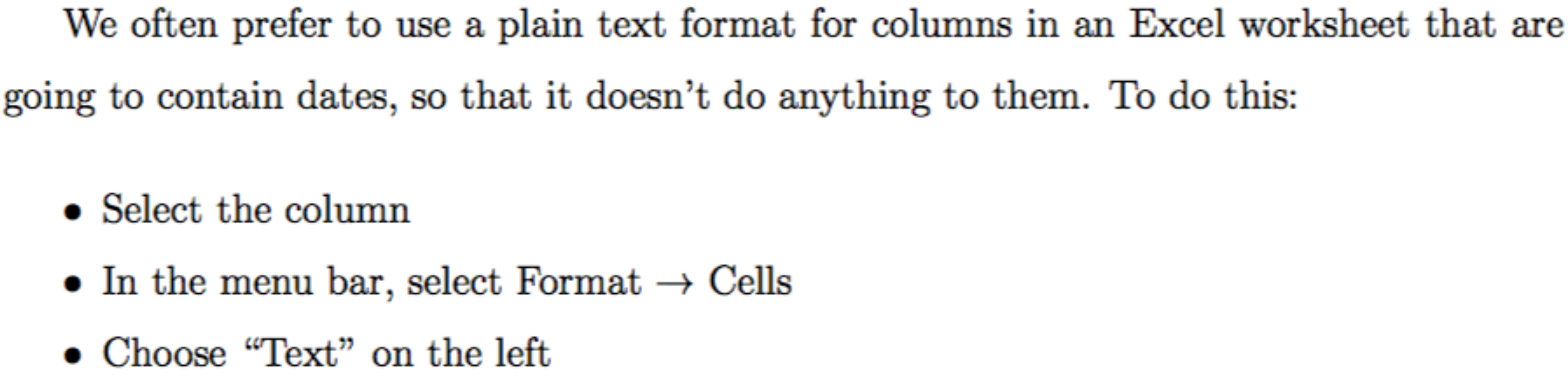

FINAL.doc!
FINAL_rev.2.doc
FINAL_rev.6.COMMENTS.doc
FINAL_rev.8.comments5.CORRECTIONS.doc
track changes
FINAL_rev.18.comments7.corrections9.MORE.30.doc
FINAL_rev.22.comments49.corrections.10.#@$%WHYDIDICOMETOGRADSCHOOL????.doc

JORGE CHAM ©2012

Write dates as YYYY-MM-DD


PUBLIC SERVICE ANNOUNCEMENT:
OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.
THIS IS THE CORRECT WAY TO WRITE NUMERIC DATES:
2013-02-27
THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:
02/27/2013 02/27/13 27/02/2013 27/02/13
20130227 2013.02.27 27.02.13 27-02-13
27.2.13 2013. II. 27. 27/2-13 2013.158904109
MMXIII-II-XXVII MMXIII LVII/CCCLXV 1330300800
((3+3)×(111+1)-1)×3/3-1/3³ 2013
10/11011/1101 02/27/20/13 01237
2 3 1 4
5 67 8
HISSSSS
2-27-13

https://xkcd.com/1179/

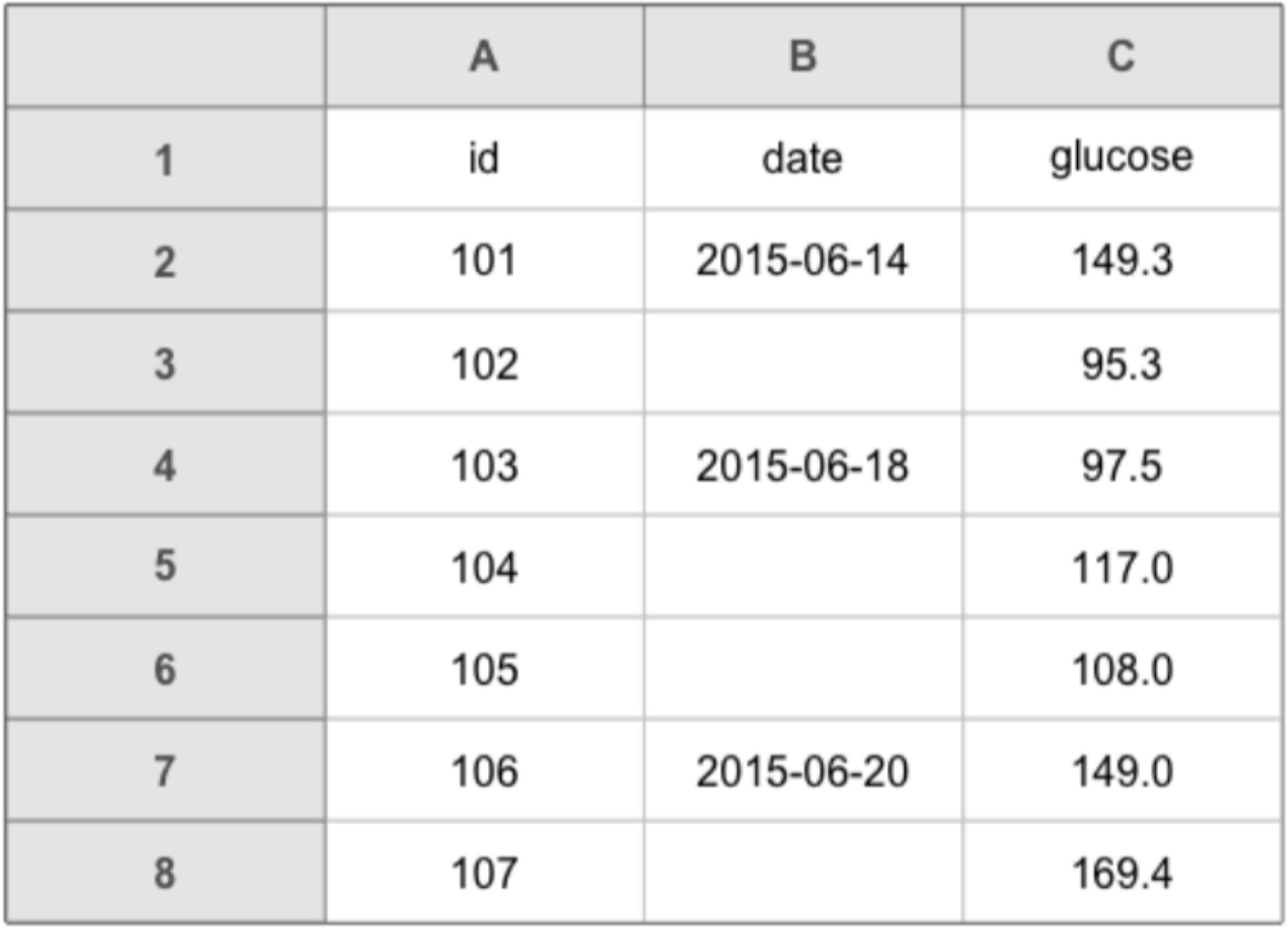


We often prefer to use a plain text format for columns in an Excel worksheet that are going to contain dates, so that it doesn't do anything to them. To do this:

- Select the column
- In the menu bar, select Format → Cells
- Choose "Text" on the left

Broman and Woo 2018

No empty cells

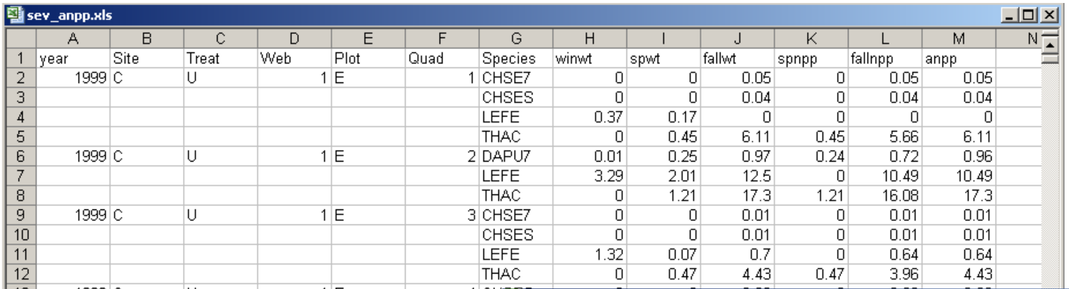


	A	B	C
1	id	date	glucose
2	101	2015-06-14	149.3
3	102		95.3
4	103	2015-06-18	97.5
5	104		117.0
6	105		108.0
7	106	2015-06-20	149.0
8	107		169.4

Broman and Woo 2018

No empty cells

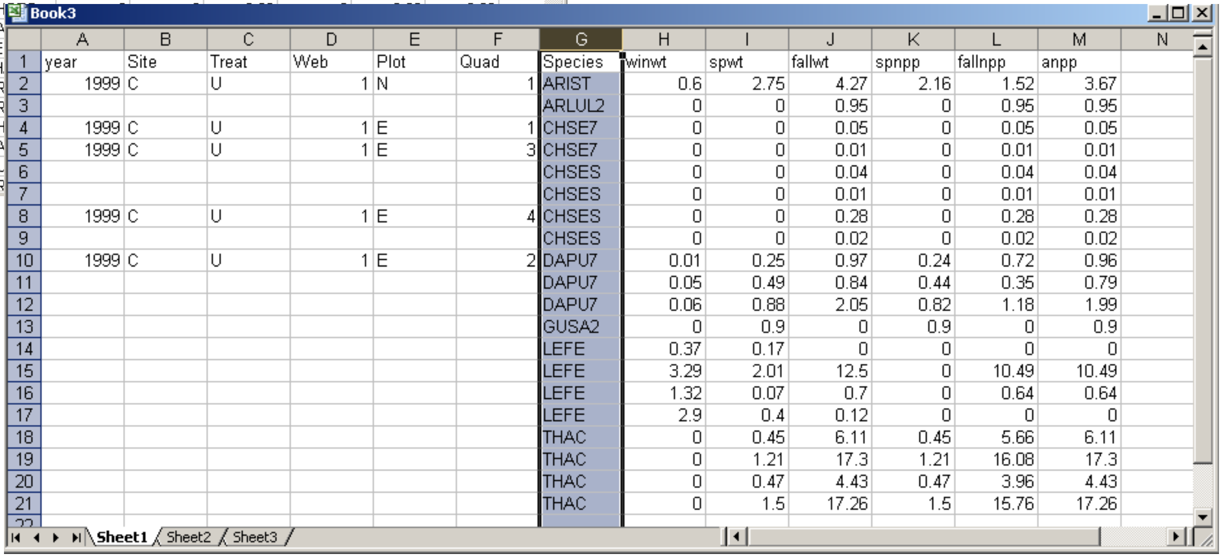
- Enter complete lines of data



sev_anpp.xls

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	year	Site	Treat	Web	Plot	Quad	Species	winwt	spwt	fallwt	spnpp	fallnpp	anpp
2	1999	C	U	1	E	1	CHSE7	0	0	0.05	0	0.05	0.05
3							CHSES	0	0	0.04	0	0.04	0.04
4							LEFE	0.37	0.17	0	0	0	0
5							THAC	0	0.45	6.11	0.45	5.66	6.11
6	1999	C	U	1	E	2	DAPU7	0.01	0.25	0.97	0.24	0.72	0.96
7							LEFE	3.29	2.01	12.5	0	10.49	10.49
8							THAC	0	1.21	17.3	1.21	16.08	17.3
9	1999	C	U	1	E	3	CHSE7	0	0	0.01	0	0.01	0.01
10							CHSES	0	0	0.01	0	0.01	0.01
11							LEFE	1.32	0.07	0.7	0	0.64	0.64
12							THAC	0	0.47	4.43	0.47	3.96	4.43
13	1999	C	U	1	E	4	CH						
14							DA						
15							LE						
16							TH						
17	1999	C	U	1	N	1	AR						
18							AR						
19							CH						
20							DA						
21							GU						
22	1999	C	U	1	N	2	AR						

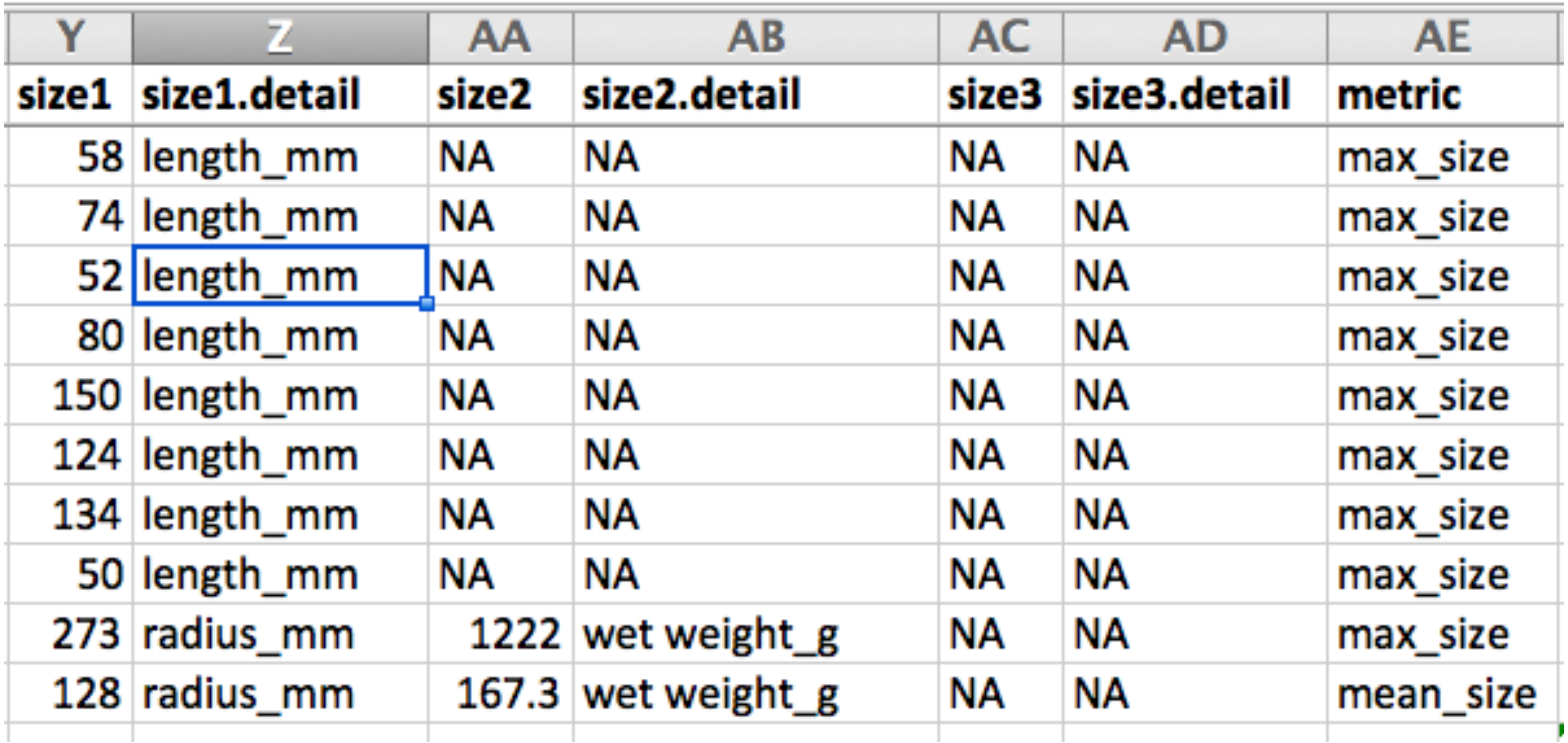
Sorting an Excel file with empty cells is not a good idea!



Book3

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	year	Site	Treat	Web	Plot	Quad	Species	winwt	spwt	fallwt	spnpp	fallnpp	anpp
2	1999	C	U	1	N	1	ARIST	0.6	2.75	4.27	2.16	1.52	3.67
3							ARLUL2	0	0	0.95	0	0.95	0.95
4	1999	C	U	1	E	1	CHSE7	0	0	0.05	0	0.05	0.05
5	1999	C	U	1	E	3	CHSE7	0	0	0.01	0	0.01	0.01
6							CHSES	0	0	0.04	0	0.04	0.04
7							CHSES	0	0	0.01	0	0.01	0.01
8	1999	C	U	1	E	4	CHSES	0	0	0.28	0	0.28	0.28
9							CHSES	0	0	0.02	0	0.02	0.02
10	1999	C	U	1	E	2	DAPU7	0.01	0.25	0.97	0.24	0.72	0.96
11							DAPU7	0.05	0.49	0.84	0.44	0.35	0.79
12							DAPU7	0.06	0.88	2.05	0.82	1.18	1.99
13							GUSA2	0	0.9	0	0.9	0	0.9
14							LEFE	0.37	0.17	0	0	0	0
15							LEFE	3.29	2.01	12.5	0	10.49	10.49
16							LEFE	1.32	0.07	0.7	0	0.64	0.64
17							LEFE	2.9	0.4	0.12	0	0	0
18							THAC	0	0.45	6.11	0.45	5.66	6.11
19							THAC	0	1.21	17.3	1.21	16.08	17.3
20							THAC	0	0.47	4.43	0.47	3.96	4.43
21							THAC	0	1.5	17.26	1.5	15.76	17.26

Just one thing in a cell



Y	Z	AA	AB	AC	AD	AE
size1	size1.detail	size2	size2.detail	size3	size3.detail	metric
58	length_mm	NA	NA	NA	NA	max_size
74	length_mm	NA	NA	NA	NA	max_size
52	length_mm	NA	NA	NA	NA	max_size
80	length_mm	NA	NA	NA	NA	max_size
150	length_mm	NA	NA	NA	NA	max_size
124	length_mm	NA	NA	NA	NA	max_size
134	length_mm	NA	NA	NA	NA	max_size
50	length_mm	NA	NA	NA	NA	max_size
273	radius_mm	1222	wet weight_g	NA	NA	max_size
128	radius_mm	167.3	wet weight_g	NA	NA	mean_size

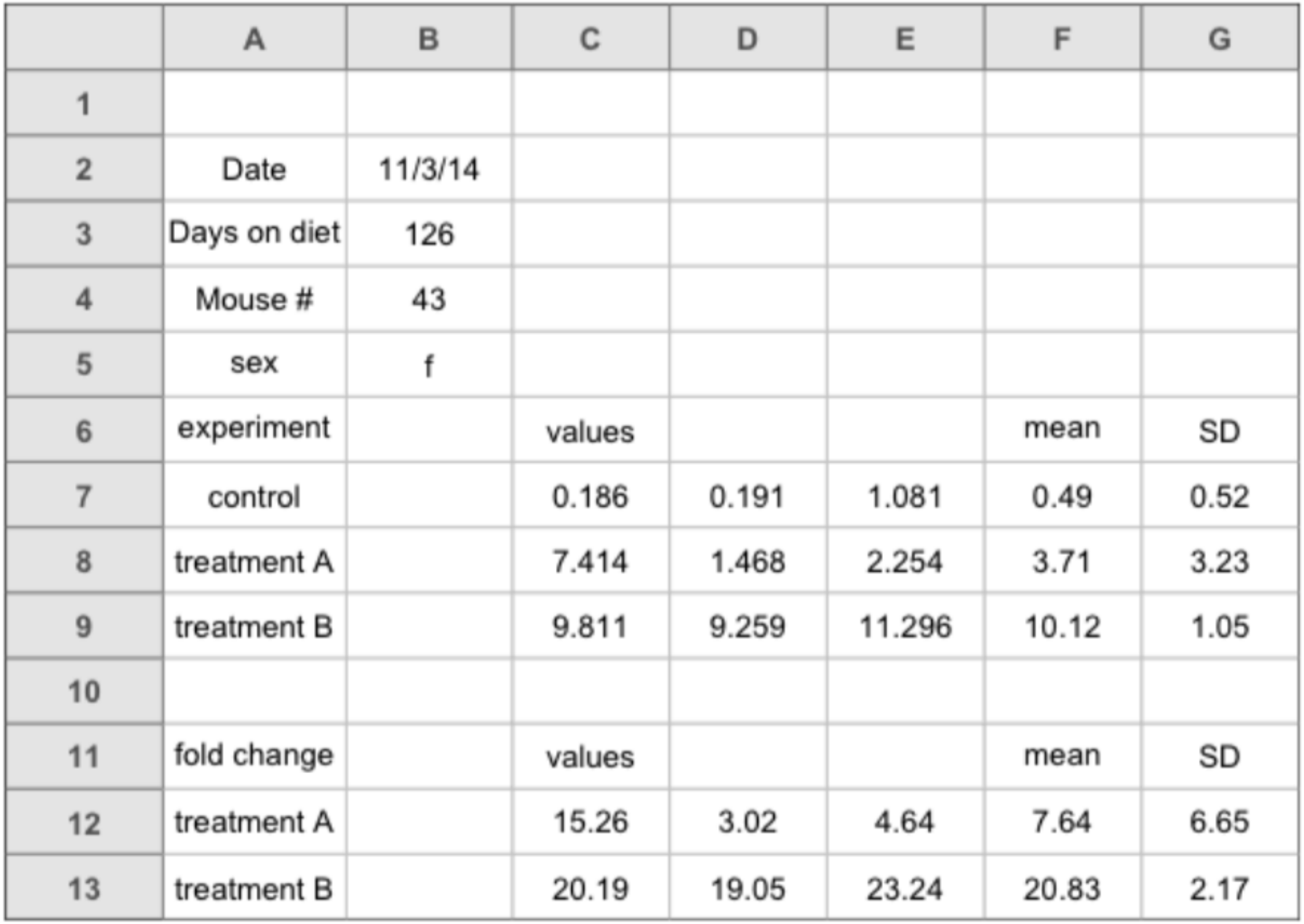
Broman and Woo 2018

Make it a rectangle

Organize the data as a single rectangle, or set of rectangles

Broman and Woo 2018

Not a rectangle



	A	B	C	D	E	F	G
1							
2	Date	11/3/14					
3	Days on diet	126					
4	Mouse #	43					
5	sex	f					
6	experiment		values			mean	SD
7	control		0.186	0.191	1.081	0.49	0.52
8	treatment A		7.414	1.468	2.254	3.71	3.23
9	treatment B		9.811	9.259	11.296	10.12	1.05
10							
11	fold change		values			mean	SD
12	treatment A		15.26	3.02	4.64	7.64	6.65
13	treatment B		20.19	19.05	23.24	20.83	2.17

Broman and Woo 2018

Make it a *tidy* rectangle

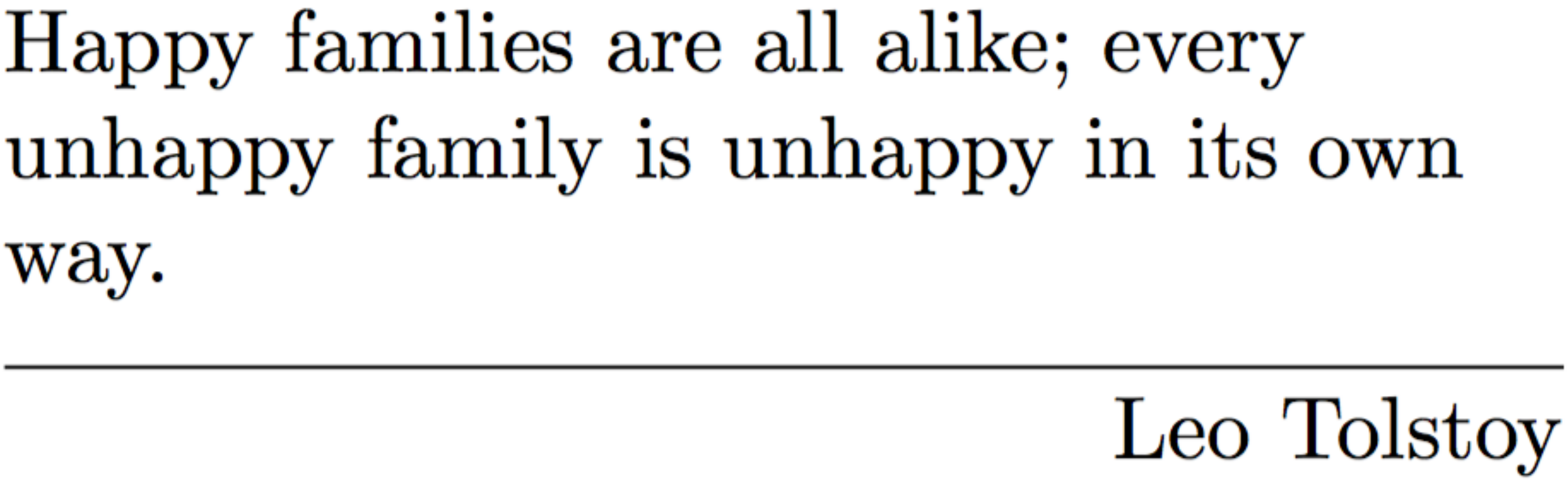
Organize the data as a single rectangle, or set of rectangles

‘Tidy’, or long, format

- subjects as rows
- variables as columns
- single header row

Broman and Woo 2018

Make it a *tidy* rectangle



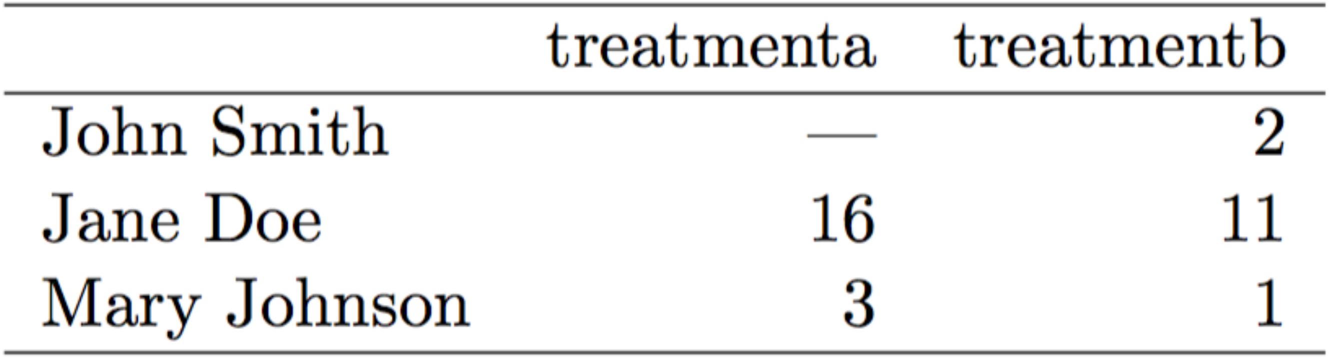
Happy families are all alike; every unhappy family is unhappy in its own way.

Leo Tolstoy

Tidy datasets are all alike; every messy dataset is messy in its own way

http://vita.had.co.nz/papers/tidy-data.html

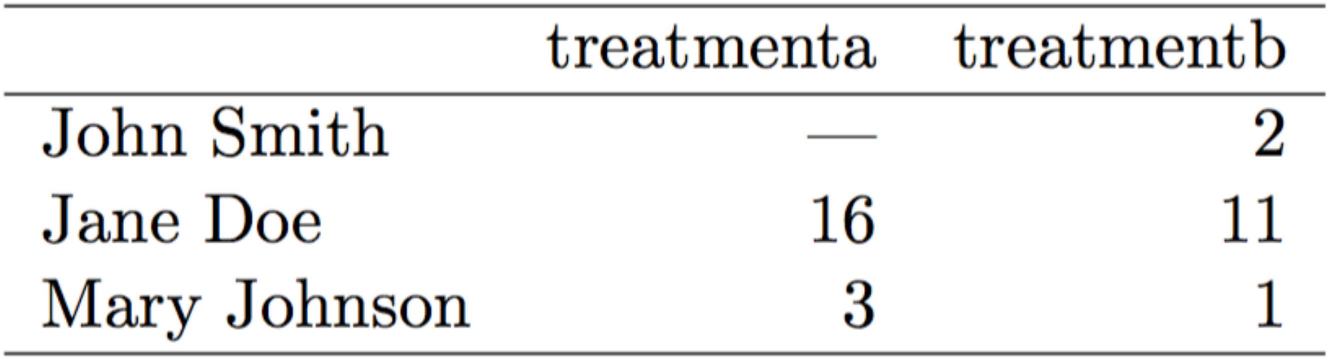
Make it a *tidy* rectangle



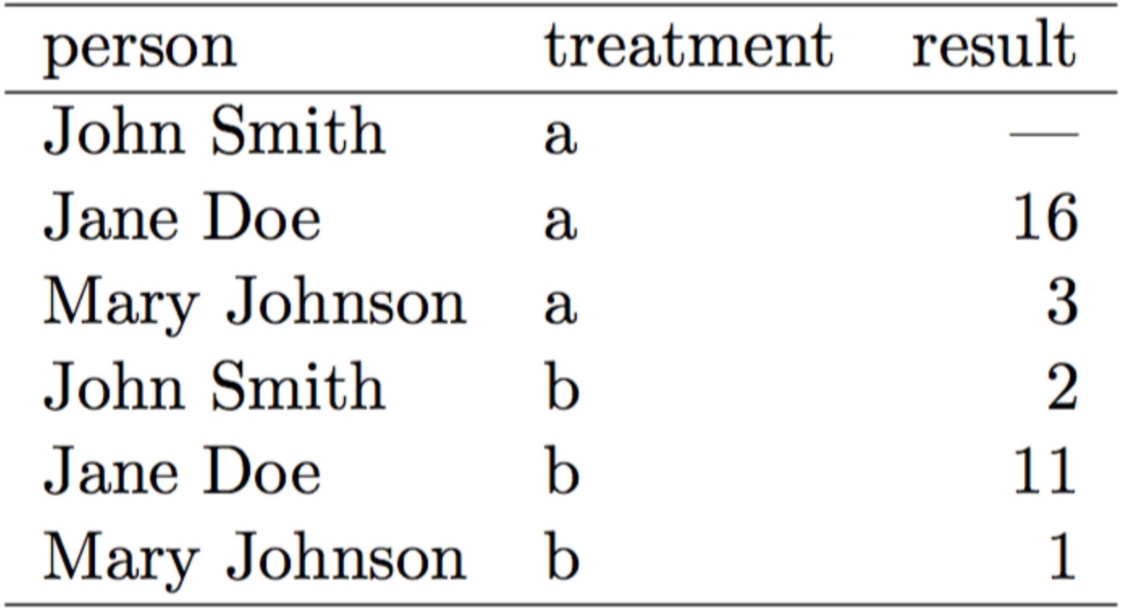
	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1

http://vita.had.co.nz/papers/tidy-data.html

Make it a *tidy* rectangle



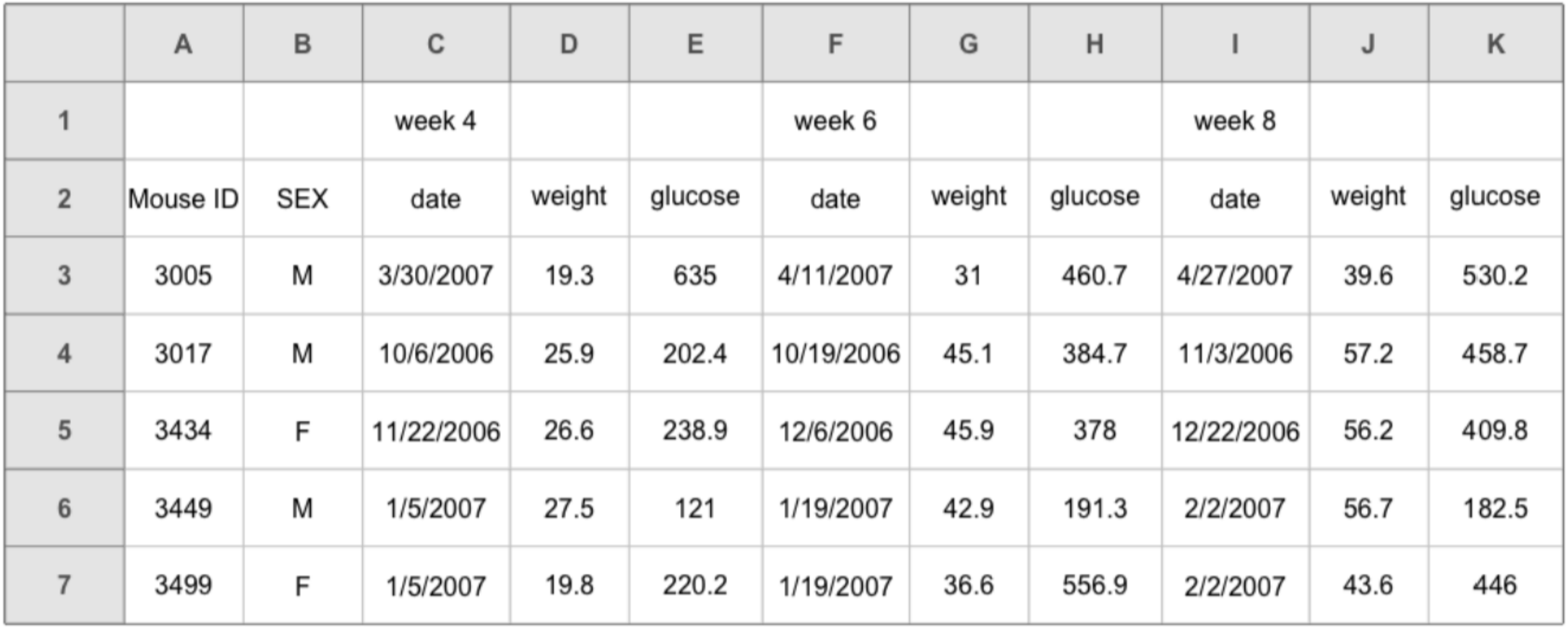
	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1



person	treatment	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

http://vita.had.co.nz/papers/tidy-data.html

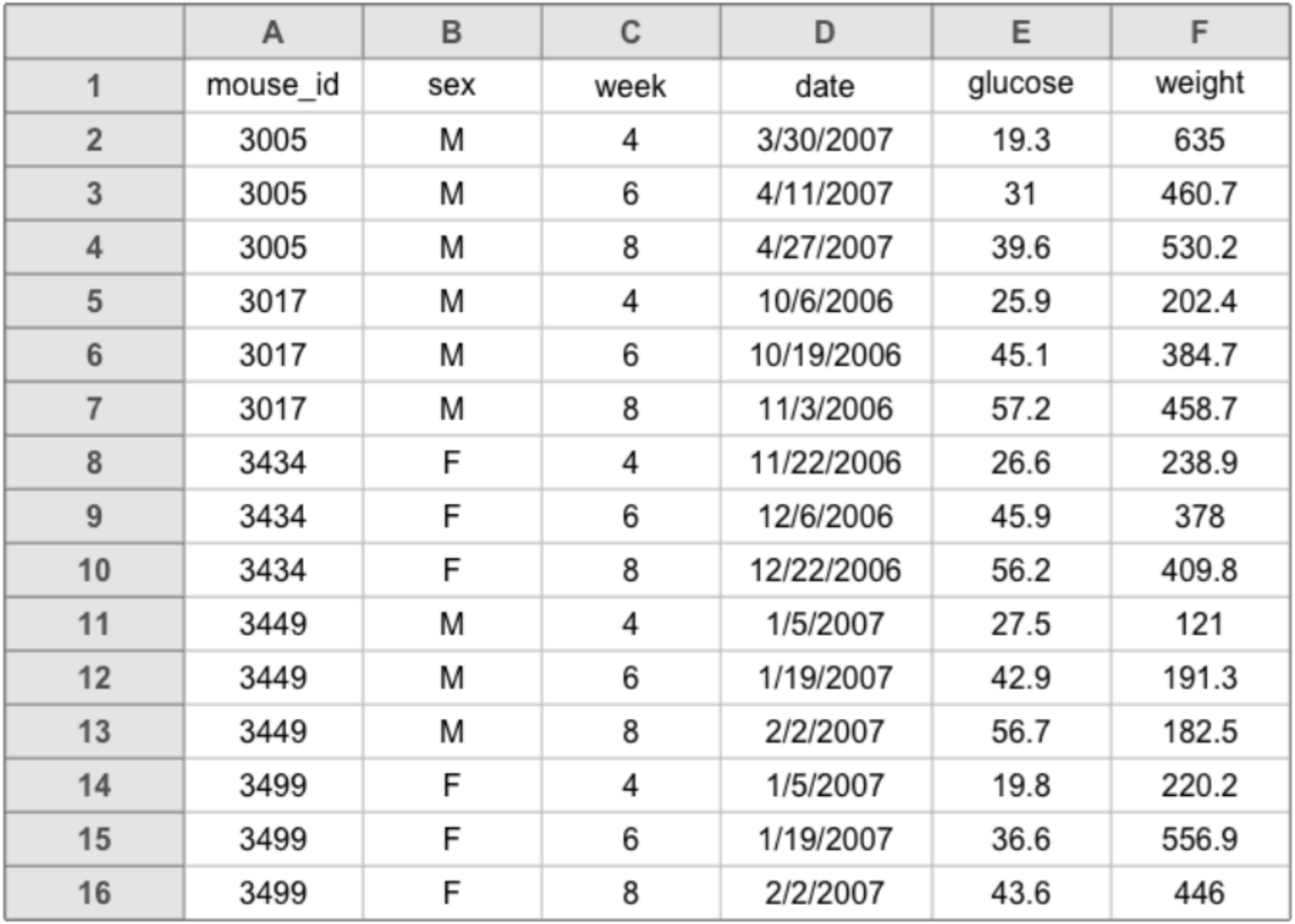
Messy



	A	B	C	D	E	F	G	H	I	J	K
1			week 4			week 6			week 8		
2	Mouse ID	SEX	date	weight	glucose	date	weight	glucose	date	weight	glucose
3	3005	M	3/30/2007	19.3	635	4/11/2007	31	460.7	4/27/2007	39.6	530.2
4	3017	M	10/6/2006	25.9	202.4	10/19/2006	45.1	384.7	11/3/2006	57.2	458.7
5	3434	F	11/22/2006	26.6	238.9	12/6/2006	45.9	378	12/22/2006	56.2	409.8
6	3449	M	1/5/2007	27.5	121	1/19/2007	42.9	191.3	2/2/2007	56.7	182.5
7	3499	F	1/5/2007	19.8	220.2	1/19/2007	36.6	556.9	2/2/2007	43.6	446

Broman and Woo 2018

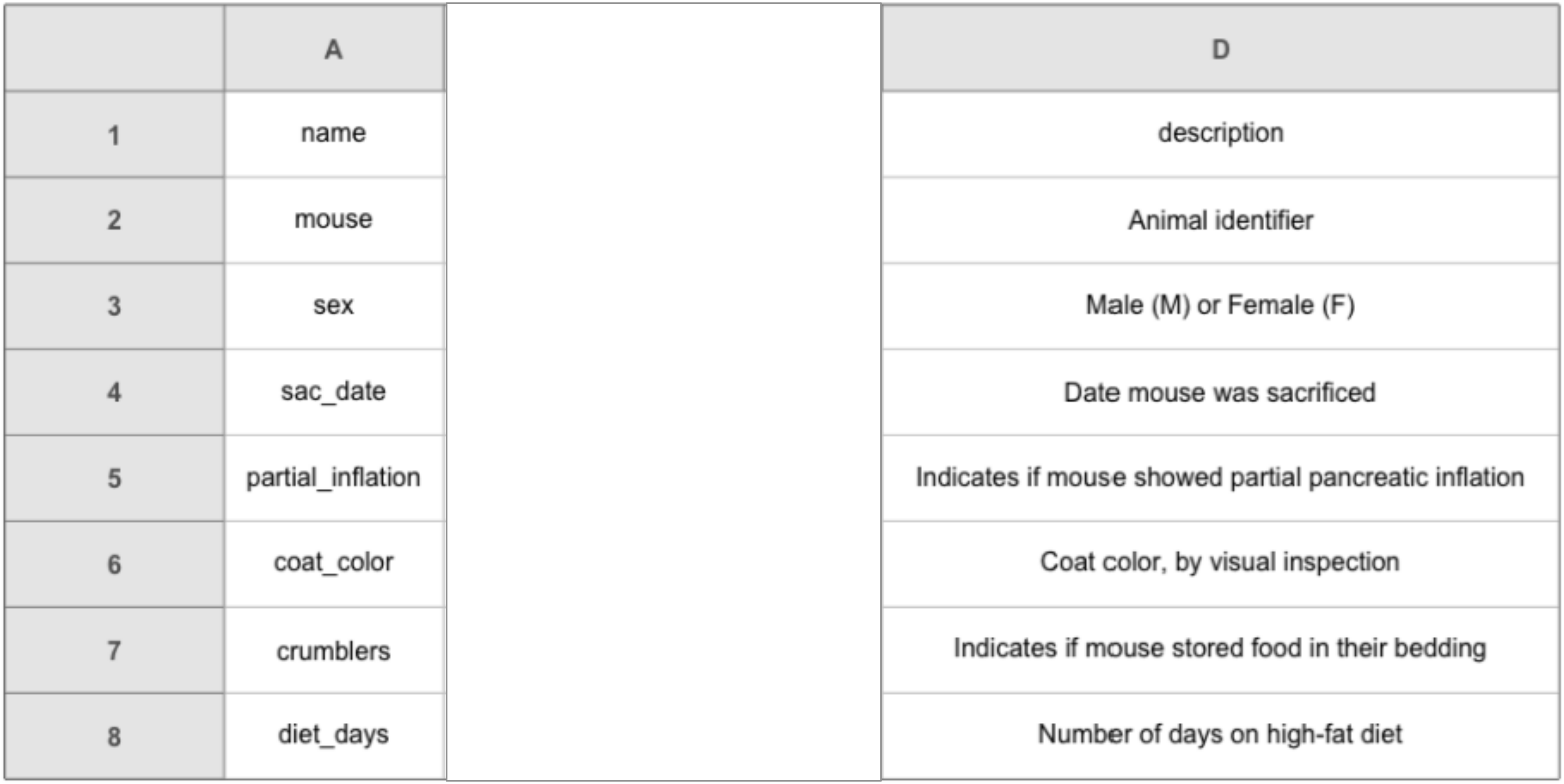
Tidy



	A	B	C	D	E	F
1	mouse_id	sex	week	date	glucose	weight
2	3005	M	4	3/30/2007	19.3	635
3	3005	M	6	4/11/2007	31	460.7
4	3005	M	8	4/27/2007	39.6	530.2
5	3017	M	4	10/6/2006	25.9	202.4
6	3017	M	6	10/19/2006	45.1	384.7
7	3017	M	8	11/3/2006	57.2	458.7
8	3434	F	4	11/22/2006	26.6	238.9
9	3434	F	6	12/6/2006	45.9	378
10	3434	F	8	12/22/2006	56.2	409.8
11	3449	M	4	1/5/2007	27.5	121
12	3449	M	6	1/19/2007	42.9	191.3
13	3449	M	8	2/2/2007	56.7	182.5
14	3499	F	4	1/5/2007	19.8	220.2
15	3499	F	6	1/19/2007	36.6	556.9
16	3499	F	8	2/2/2007	43.6	446

Broman and Woo 2018

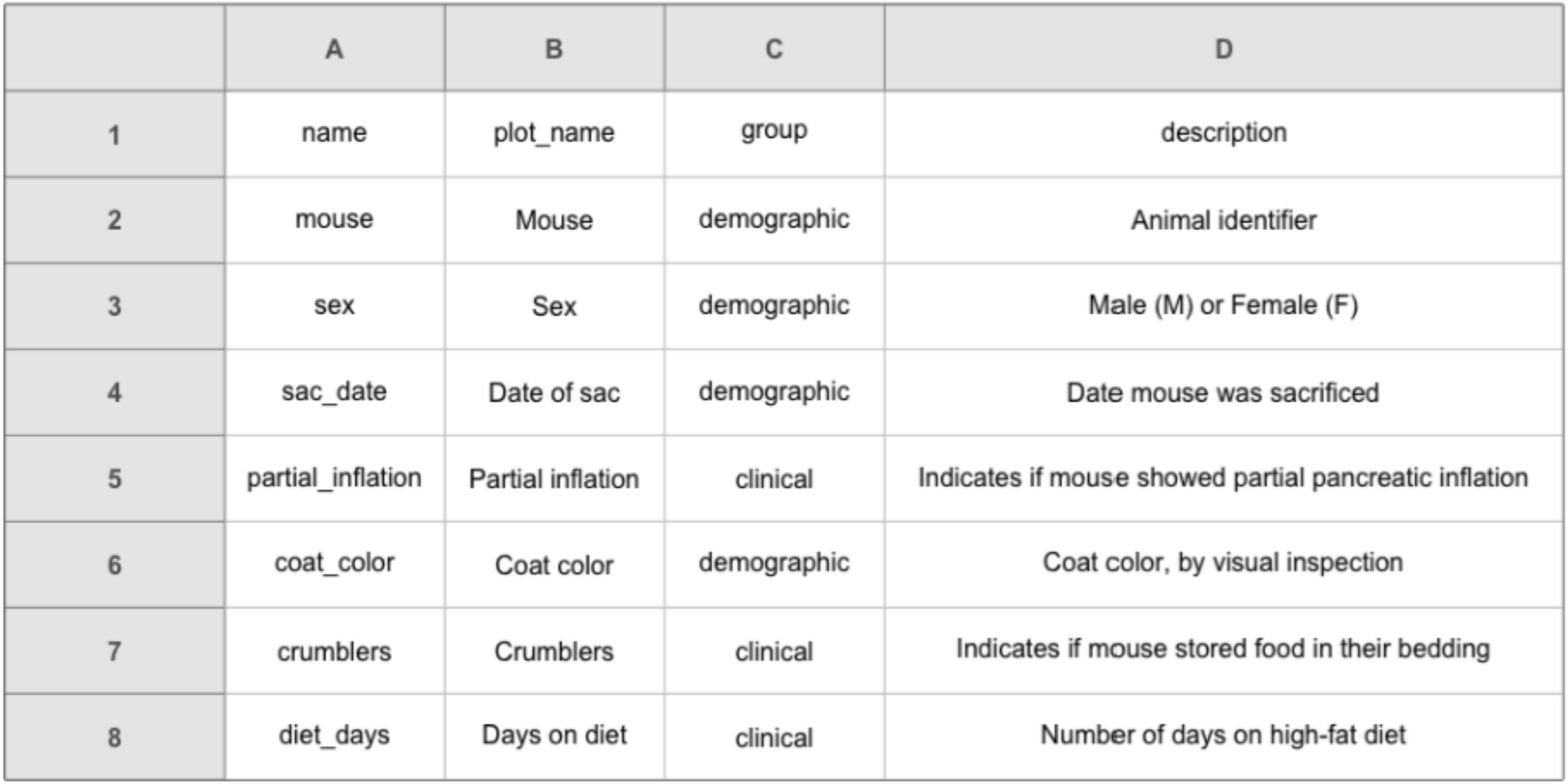
Create a data dictionary



	A	D
1	name	description
2	mouse	Animal identifier
3	sex	Male (M) or Female (F)
4	sac_date	Date mouse was sacrificed
5	partial_inflation	Indicates if mouse showed partial pancreatic inflation
6	coat_color	Coat color, by visual inspection
7	crumblers	Indicates if mouse stored food in their bedding
8	diet_days	Number of days on high-fat diet

Broman and Woo 2018

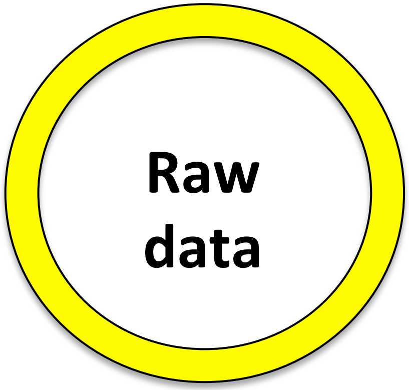
Create a data dictionary



	A	B	C	D
1	name	plot_name	group	description
2	mouse	Mouse	demographic	Animal identifier
3	sex	Sex	demographic	Male (M) or Female (F)
4	sac_date	Date of sac	demographic	Date mouse was sacrificed
5	partial_inflation	Partial inflation	clinical	Indicates if mouse showed partial pancreatic inflation
6	coat_color	Coat color	demographic	Coat color, by visual inspection
7	crumblers	Crumblers	clinical	Indicates if mouse stored food in their bedding
8	diet_days	Days on diet	clinical	Number of days on high-fat diet

Broman and Woo 2018

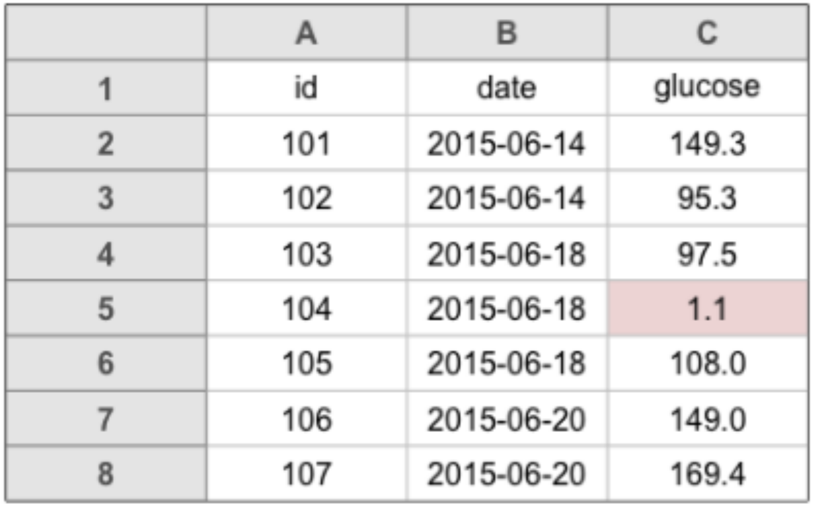
No calculations



Raw data

Broman and Woo 2018

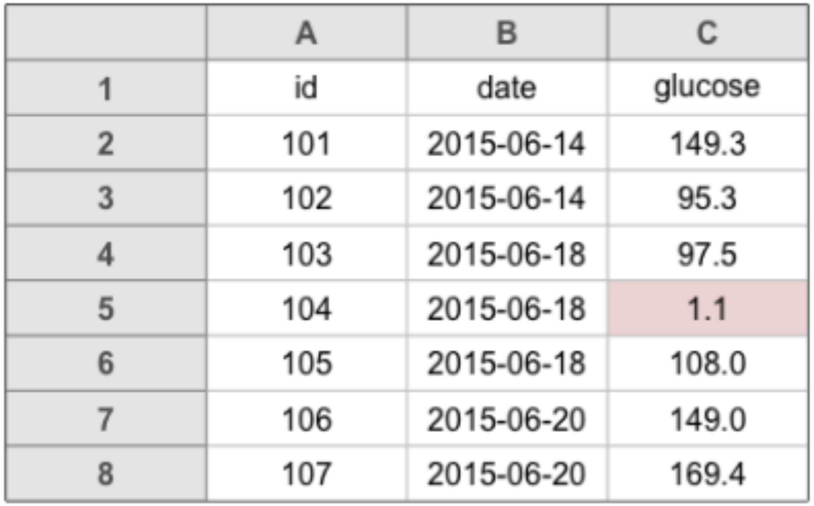
Don’t use font color or highlighting as data



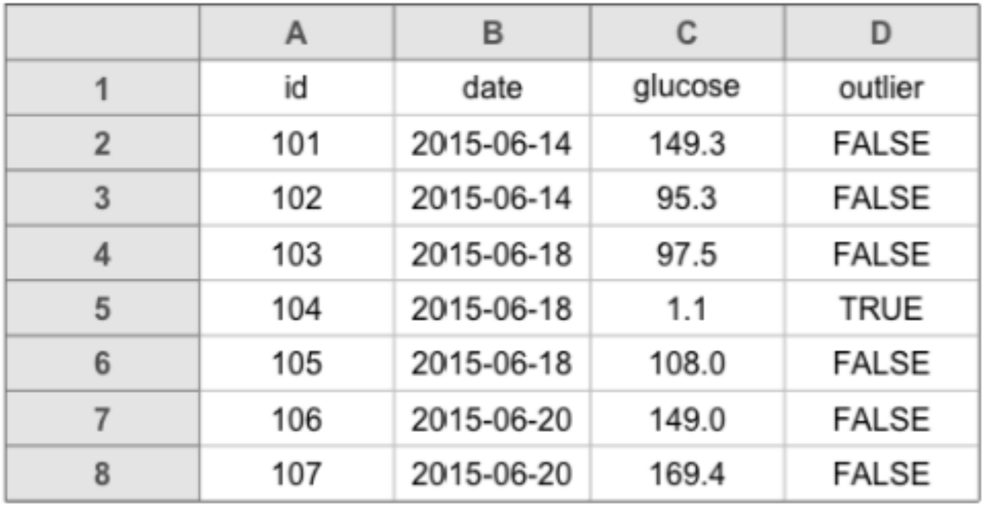
	A	B	C
1	id	date	glucose
2	101	2015-06-14	149.3
3	102	2015-06-14	95.3
4	103	2015-06-18	97.5
5	104	2015-06-18	1.1
6	105	2015-06-18	108.0
7	106	2015-06-20	149.0
8	107	2015-06-20	169.4

Broman and Woo 2018

Don't use font color or highlighting as data



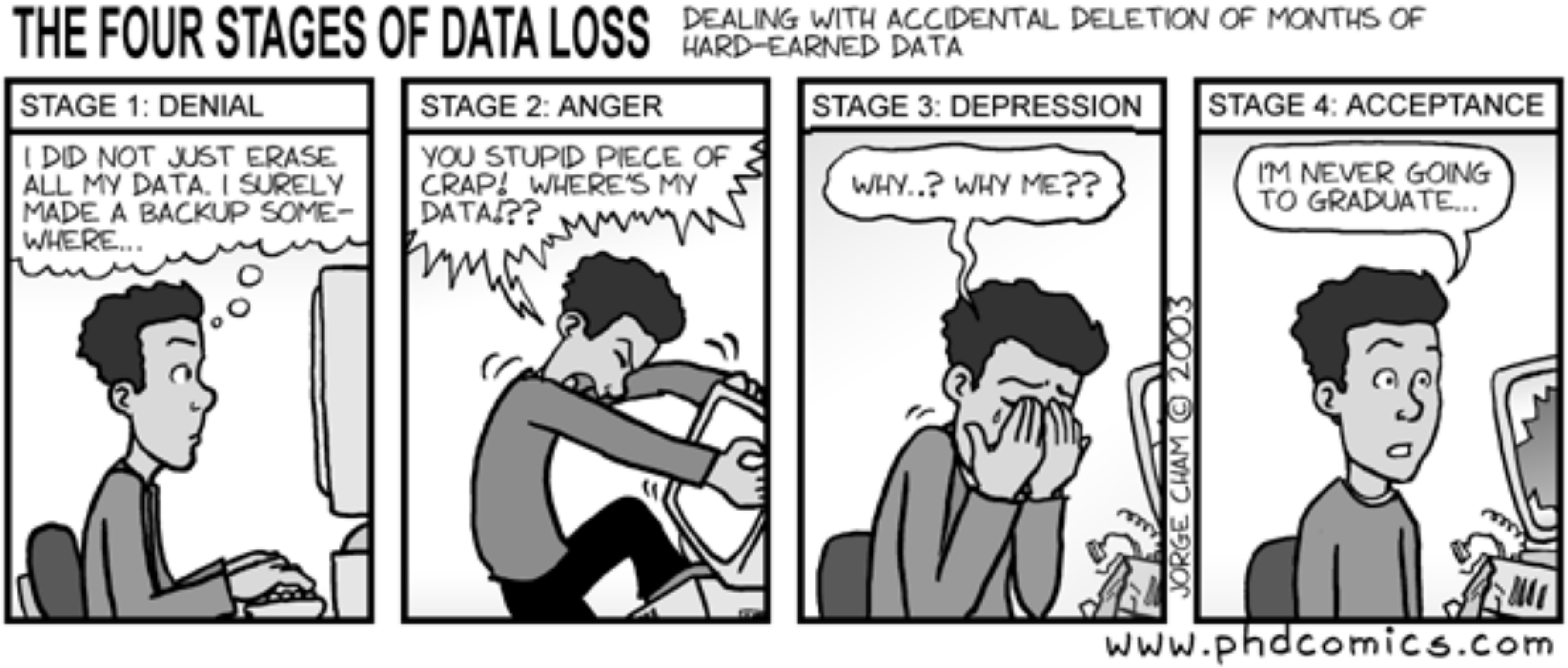
	A	B	C
1	id	date	glucose
2	101	2015-06-14	149.3
3	102	2015-06-14	95.3
4	103	2015-06-18	97.5
5	104	2015-06-18	1.1
6	105	2015-06-18	108.0
7	106	2015-06-20	149.0
8	107	2015-06-20	169.4



	A	B	C	D
1	id	date	glucose	outlier
2	101	2015-06-14	149.3	FALSE
3	102	2015-06-14	95.3	FALSE
4	103	2015-06-18	97.5	FALSE
5	104	2015-06-18	1.1	TRUE
6	105	2015-06-18	108.0	FALSE
7	106	2015-06-20	149.0	FALSE
8	107	2015-06-20	169.4	FALSE

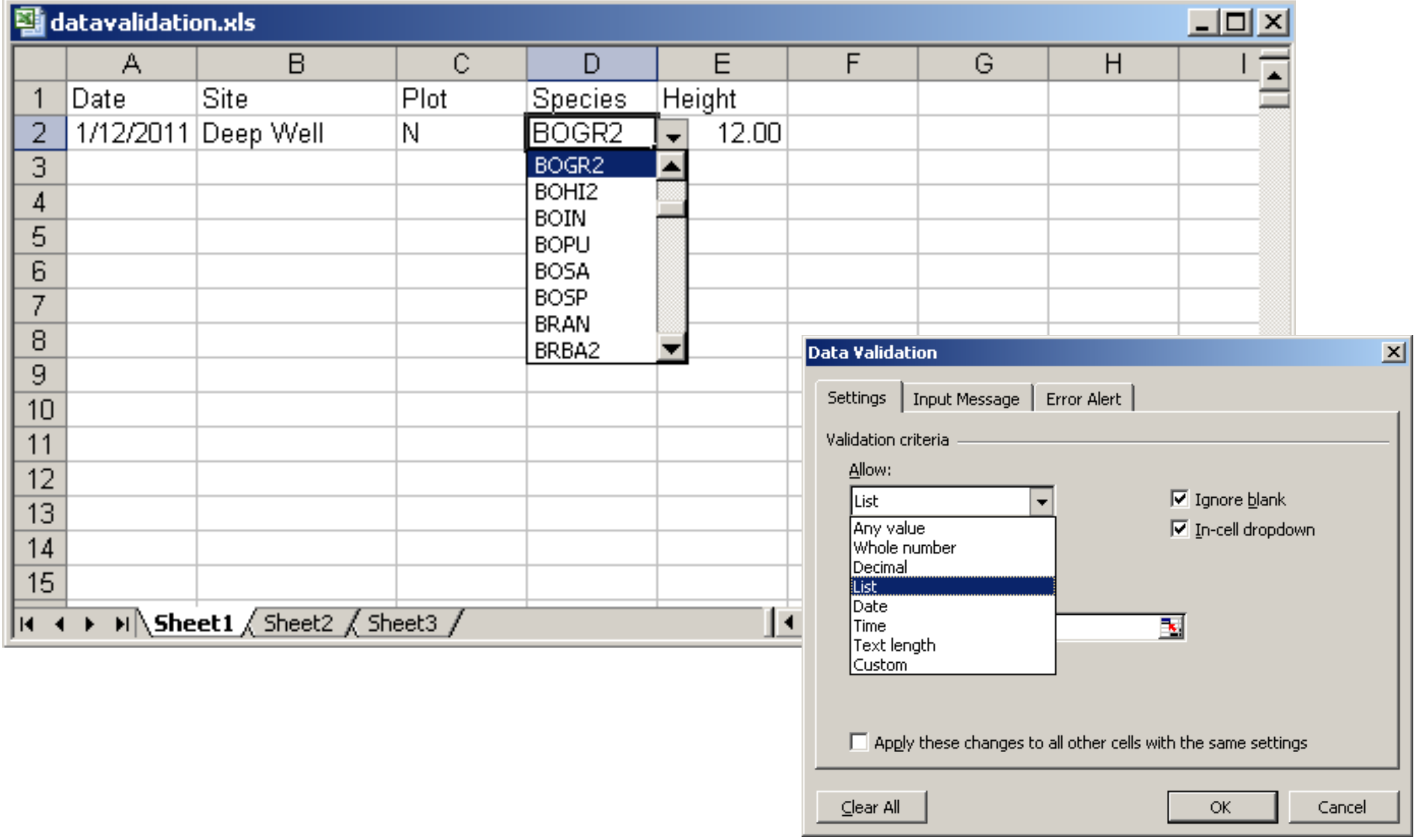
Broman and Woo 2018

Make backups


THE FOUR STAGES OF DATA LOSS
DEALING WITH ACCIDENTAL DELETION OF MONTHS OF HARD-EARNED DATA
STAGE 1: DENIAL
I DID NOT JUST ERASE ALL MY DATA. I SURELY MADE A BACKUP SOMEWHERE...
STAGE 2: ANGER
YOU STUPID PIECE OF CRAP! WHERE'S MY DATA!??
STAGE 3: DEPRESSION
WHY..? WHY ME??
STAGE 4: ACCEPTANCE
I'M NEVER GOING TO GRADUATE...
JORGE CHAM © 2003
www.phdcomics.com

Broman and Woo 2018

Use data validation to avoid data entry errors



datavalidation.xls

	A	B	C	D	E
1	Date	Site	Plot	Species	Height
2	1/12/2011	Deep Well	N	BOGR2	12.00

BOGR2
BOHI2
BOIN
BOPU
BOSA
BOSP
BRAN
BRBA2

Data Validation

Settings | Input Message | Error Alert

Validation criteria

Allow: List

Any value
Whole number
Decimal
List
Date
Time
Text length
Custom

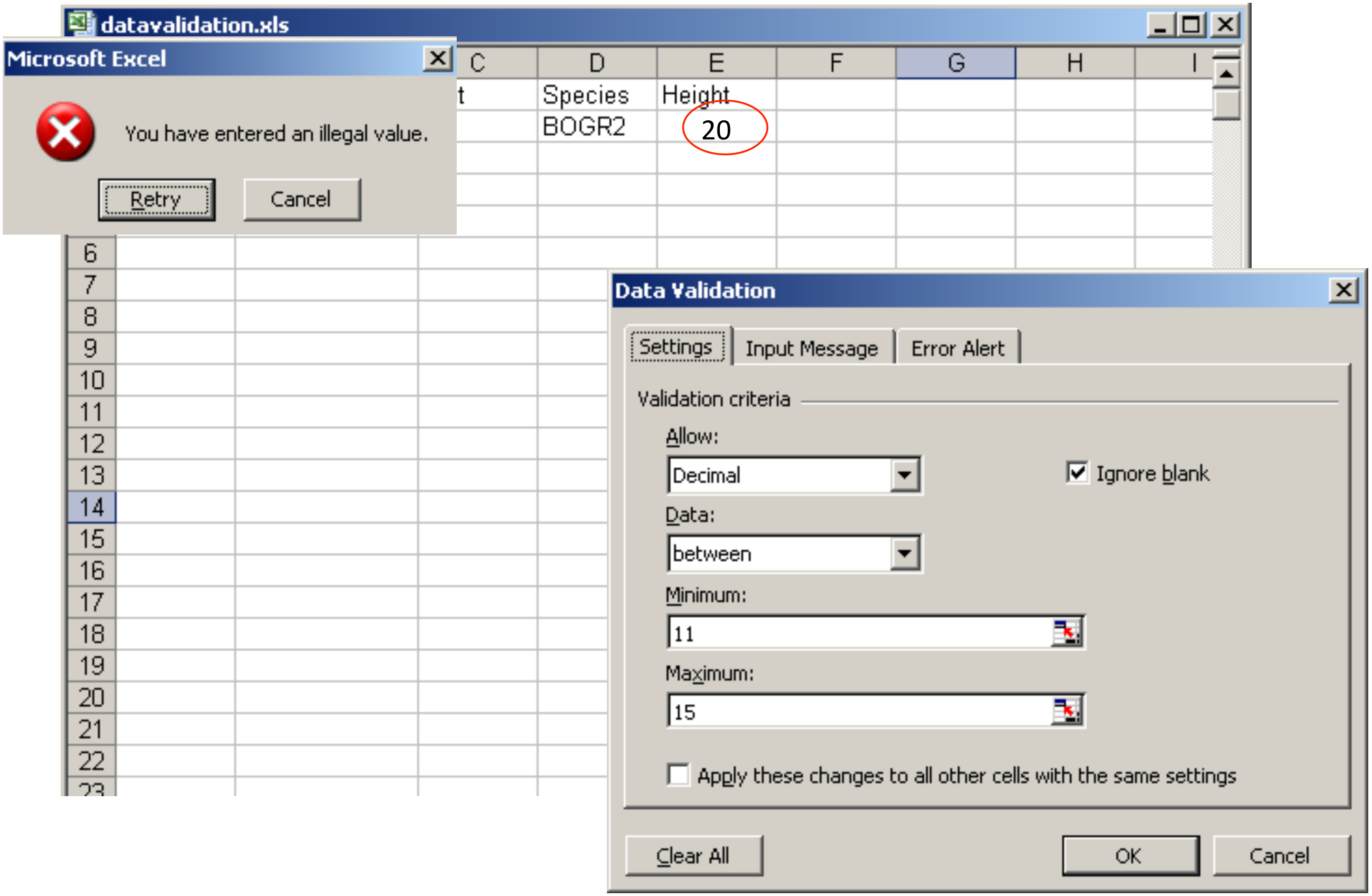
☑ Ignore blank
☑ In-cell dropdown

☐ Apply these changes to all other cells with the same settings

Clear All | OK | Cancel

Broman and Woo 2018

Use data validation to avoid data entry errors



datavalidation.xls

C	D	E
t	Species	Height
	BOGR2	20

Microsoft Excel

You have entered an illegal value.

Data Validation

Validation criteria

Allow: Decimal

☑ Ignore blank

Data: between

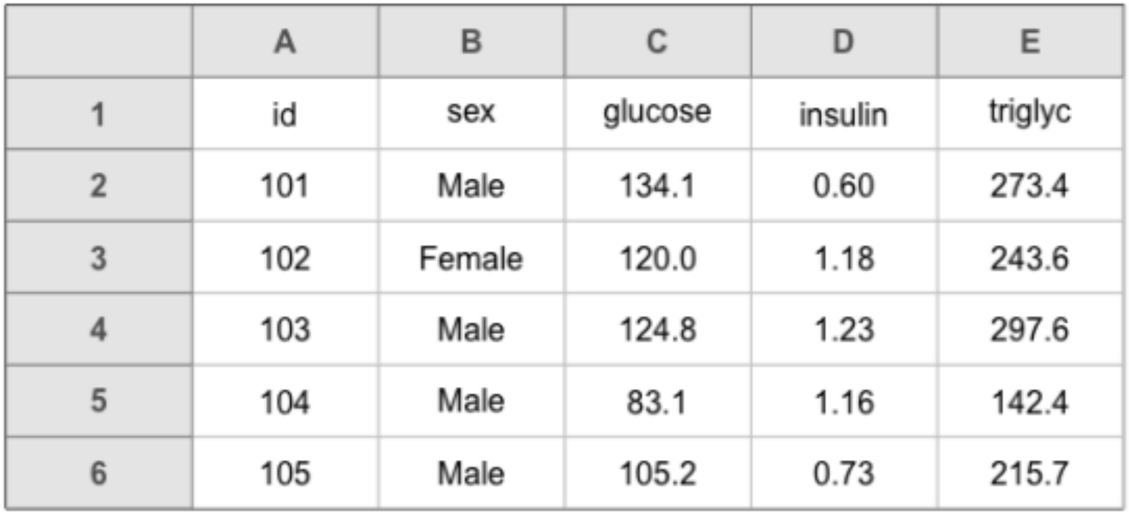
Minimum: 11

Maximum: 15

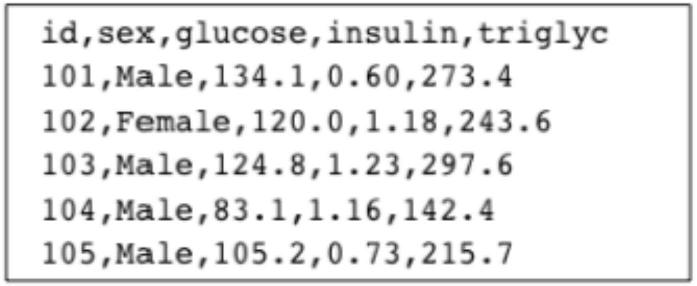
□ Apply these changes to all other cells with the same settings

Broman and Woo 2018

Save the data in plain text files



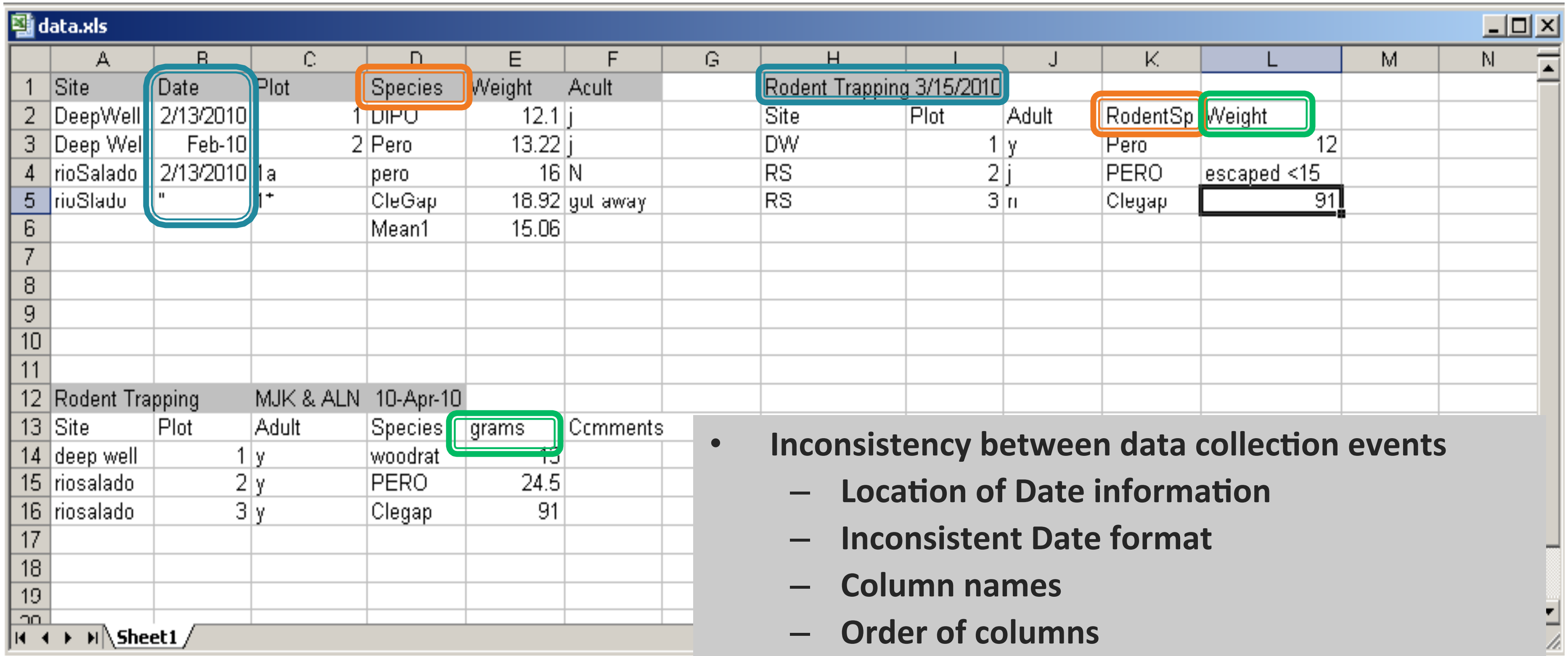
	A	B	C	D	E
1	id	sex	glucose	insulin	triglyc
2	101	Male	134.1	0.60	273.4
3	102	Female	120.0	1.18	243.6
4	103	Male	124.8	1.23	297.6
5	104	Male	83.1	1.16	142.4
6	105	Male	105.2	0.73	215.7



```
id,sex,glucose,insulin,triglyc
101,Male,134.1,0.60,273.4
102,Female,120.0,1.18,243.6
103,Male,124.8,1.23,297.6
104,Male,83.1,1.16,142.4
105,Male,105.2,0.73,215.7
```

Broman and Woo 2018

Example: Poor Data Entry



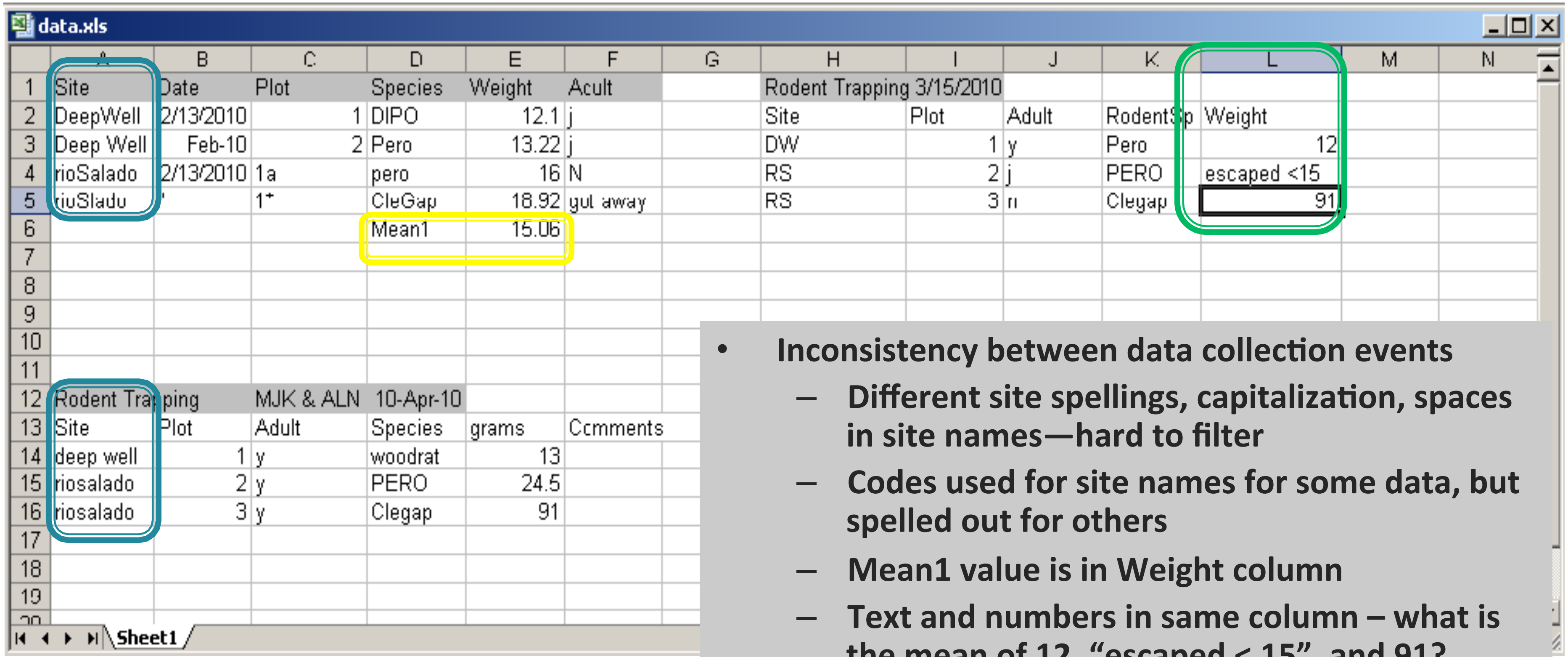
data.xls

	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	1 a	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	15									
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

DataONE

Example: Poor Data Entry

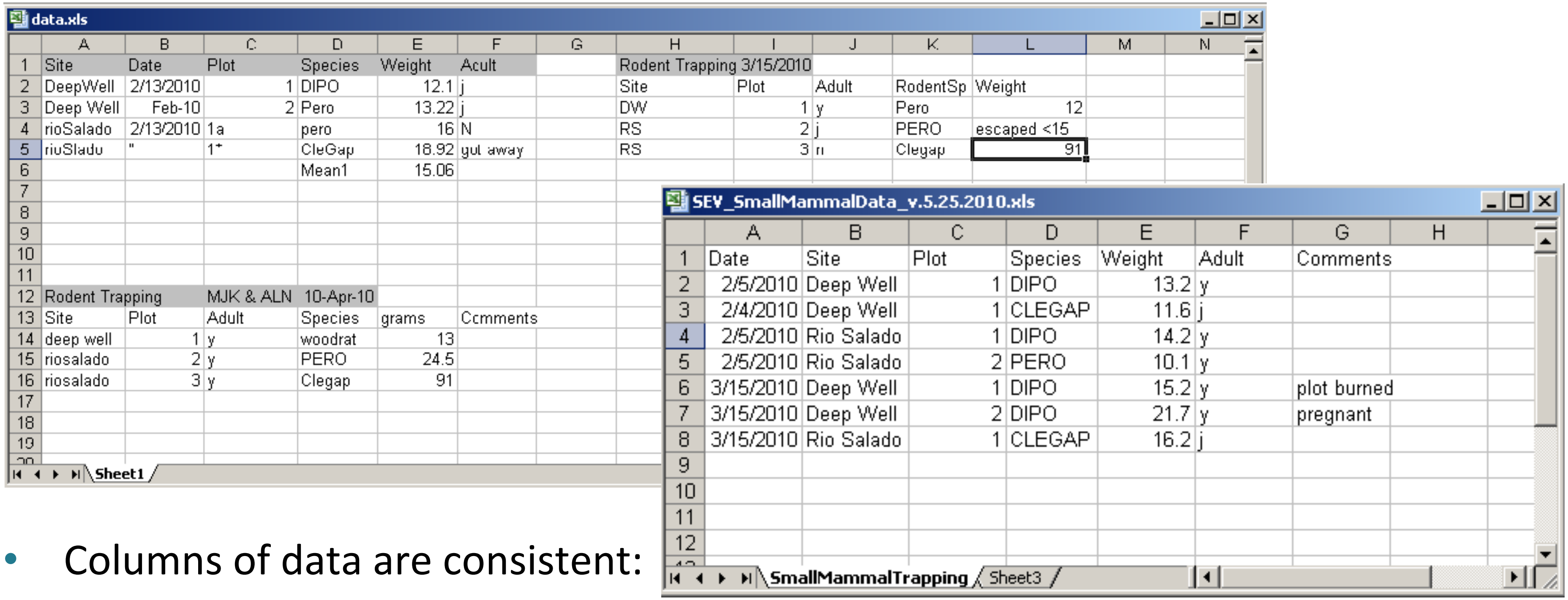


data.xls

	A	B	C	D	E	F	G	H	I	J	K	L
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 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							
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							

- 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?

Best Practices



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 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														

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		
9								
10								
11								
12								

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

Class exercise

You should have 3 (fictional) data files: pond2010.xlsx, zoop-temp-main.xlsx; zoop-temp.xlsx.

These 3 files were all intended to be part of the same study – the investigators wanted to examine the day-night distribution of 2 species of zooplankton across multiple years. The type of zooplankton they studied is called rotifers generally, and specifically the genus *Conochilus,* in which groups of individual rotifers stick together in colonies (see http://eol.org/pages/43393/overview). The investigators plan to repeat this study for several more years.

Activity 1

As individuals or in small groups, open the 3 files and inspect them. Based on what you have learned so far about data management, what are some problems in the way the data are currently organized?

Activity 2

Suggest a new system for organization. Create a new spreadsheet that can be used as a template for later years of data collection.