How to avoid data disaster

Valeria Bo
Sergio Martinez

Data

Tables, lists, …

Numbers, names, dates, …

Graphs, images, ...

Popular file format: .txt .csv, .xls, …

Always, always **RAW** data

NOT

- Processed
- Filtered
- Manipulated

either electronically or manually

Why? Because…

Maintain consistency

- Data format, codes (M,F)
- Separator: , ; tab

Reduce human errors

- Copy / Paste / Cut
- Delete
- Addition unwanted characters

Why? Because…

Reduce machine errors

- Cell formats
- Save file with a different extension

Example 1

Patient ID	Sex	Date of birth
1	M	01-01-2013
2	f	04-18-1998
3	Male	1st April 2004
4	Female	NA
5	F	2010/03/12
6	F	
7	M	10012012

Example 1

- Consistency: F, female, f, fem, 2, …
- Single common format for all dates

  YYYY-MM-DD or YYYYMMDD

- Consistency about missing values
- NA(not available), NULL, , ...

Example 1 - modified

Patient ID	Sex	Date of birth
1	M	2013-01-01
2	F	1998-04-18
3	M	2004-04-01
4	F	NA
5	F	2010-03-12
6	F	NA
7	M	2012-01-10

A bit more about consistency

- Realistic and easy to understand
  - Variable names
  - File names
- Unique and consistent variable names
  - Multiple tables

Blank spaces

Be careful about extra spaces within cells

- Blank cell is different then a cell that contains a single space
- “male” is different from “ male ”
- Last line “ ”

These can be a headache later on!

Example 2

Patient ID	Date	Value
1	2015-06-14	123
2		76.5
3	2015-06-18	32
4		120.3
5		109
6	2015-06-20	105
7		143

Example 2

Fill in all cells

- Problems when sorting

Empty cell

- Missing value?
- Value meant to be repeated multiple times?

Missing value -> NA

Example 2

Make sure it's clear that the data is **missing**
and
not unintentionally left blank

Example 2

Patient ID	Date	Value
1	2015-06-14	123
2	2015-06-14	76.5
3	2015-06-18	32
4	2015-06-18	120.3
5	2015-06-18	109
6	2015-06-20	105
7	2015-06-20	143

Example 3

	A	B	C	D	E	F	G	H	I
1		1 min				5 min			
2	strain	normal		mutant		normal		mutant	
3	A	147	139	166	179	334	354	451	474
4	B	246	240	178	172	514	611	412	447

Example 3

	A	B	C	D	E
1	strain	genotype	min	replicate	response
2	A	normal	1	1	147
3	A	normal	1	2	139
4	B	normal	1	1	246
5	B	normal	1	2	240
6	A	mutant	1	1	166
7	A	mutant	1	2	179
8	B	mutant	1	1	178
9	B	mutant	1	2	172
10	A	normal	5	1	334
11	A	normal	5	2	354
12	B	normal	5	1	514
13	B	normal	5	2	611
14	A	mutant	5	1	451
15	A	mutant	5	2	474
16	B	mutant	5	1	412
17	B	mutant	5	2	447

No empty cells!

More…

Don’t put too much information in 1 cell

1 cell = 1 information

Don’t include units such as “30 g”.

“g” in the column name

“0 (below threshold)”

write notes in a separate column

More…

Make it **rectangle**

Create a data dictionary – separate file

Avoid using “,” or “;” or tab

Do not manually modify values–copies

No calculations

No font colour or highlighting

computer doesn’t recognize it

Good vs Bad Name

Good Name	**Good alternative**	**Bad name**
MaxTemp	max_temp	Maximum Temperature (C)
Quantity	Quantity_mg	Quamg
Sex		M/F
Weight	Weight_kg	w

Write-protect

Mac:

- Right-click on the file in Finder
- Select “Get Info”
- Sharing and permission
- Priviledge
- Read only

Write-protect

Windows:

- Right-click on the file in windows explorer
- Properties
- General tab
- Attributes
- Select the box for “read only” and click ok

Data validation

Excel data validation feature

- Select a column
- In the menu bar, choose Data
- Validation

Integer or decimal number - range

List of possible values

Limited length text

Be careful

- When identifiers are long integers
  - 1000000 -> 1e06
- Do not fill blank cells with 0s
  - 0s are data!

Save

Always keep a copy of your data files in a plain text format

-Tab delimited

-, or ; separated

Such as .csv or .txt

BACKUP!!!

Practise 1

IDs

	A
1	Trio
2	No
3	AA
4	BB/AA
5	BB/AA
6	BB/AA
7	No
8	AA
9	No
10	No
11	No
12	BB/AA
13	No
14	AA
15	BB/AA
16	FutureBB/AA
17	AA
18	FutureBB/AA
19	FutureBB/AA
20	AA
21	BB/AA
22	AA
23	BB/AA
24	

- Column titles
- Missing values

Name consistency

H	I
AAID1	AAID2
XXX193/XXX330	N/A/N/A
XXX240/XXX192/XXX138/XXX149	N/A/N/A/N/A/N/A
XXX191/XXX239	N/A/N/A
XXX243/XXX336	N/A/N/A
XXX338/XXX244	N/A/N/A
XXX194/XXX147	N/A/N/A
XXX242/XXX195/XXX141	N/A/N/A/N/A
XXX196/XXX148	N/A/N/A
XXX197/NONE/XXX122	N/A/N/A/N/A
XXX198/XXX328	N/A/N/A
XXX241/XXX199	N/A/N/A
XXX200/XXX329	N/A/N/A
XXX252/XXX226/XXX139	N/A/N/A/N/A
XXX167/XXX344	N/A/N/A
XXX327/XXX250	N/A/N/A
XXX229/XXX253/XXX140	N/A/N/A/N/A
XXX230/XXX251	N/A/N/A
XXX168/XXX347	N/A/N/A
XXX170/XXX325/XXX134	N/A/N/A/N/A
XXX209/XXX265	N/A/N/A
XXX210/XXX257/XXX128	N/A/N/A/XXX567
XXX205/XXX172	N/A/N/A

- ID1/ID2
  - –Separate columns
- N/A/N/A
  - –NA/NA
  - –Separate columns

J	K
BBID1	BBID2
LLL6005186-DNA_F04/LLL6005185-DNA_F04	/
LLL6005186-DNA_E04/LLL6005185-DNA_E04	/
LLL6005185-DNA_D04/LLL6005627-DNA_D05	/NULL
LLL6005858-DNA_B01/LLL6005628-DNA_B01	Resent due to QC failure. Old ID = LLL7-DNA_B01/NULL
LLL6005858-DNA_A01/LLL6005628-DNA_A01	Resent due to QC failure. Old ID = LLL7-DNA_A01/NULL
LLL6005186-DNA_G04/LLL6005185-DNA_G04	/
LLL6005186-DNA_H04/LLL6005185-DNA_H04	/
LLL6005186-DNA_A05/LLL6005185-DNA_A05	/
LLL6005186-DNA_B05/LLL6005185-DNA_B05	/
LLL6005186-DNA_C05/LLL6005185-DNA_C05	/
LLL6005186-DNA_E05/LLL6005185-DNA_D05	LLL-DNA_D05 => LLL-DNA_E05 due to genotype mismatch/
LLL6005186-DNA_D05/LLL6005185-DNA_E05	LLL-DNA_E05 => LLL-DNA_D05 due to genotype mismatch/
LLL6005521-DNA_H01/LLL6005520-DNA_H01	NULL/NULL
LLL6005858-DNA_C01/LLL6005628-DNA_C01	Resent due to QC failure. Old ID = LLL7-DNA_C01/NULL
NA/NA	NA/NA
LLL6005521-DNA_B02/LLL6005520-DNA_B02	NULL/NULL
NA/NA	NA/NA
NA/NA	NA/NA
NA	NA
LLL6005341-DNA_C02/LLL6005343-DNA_C02	NULL/NULL
LLL6005341-DNA_D02/LLL6005343-DNA_D02	NULL/NULL
LLL6005341-DNA_F01/LLL6005343-DNA_F01	NULL/NULL

“/” or “N/A” or “NULL”?

Practise 2