Welcome to the CRUK Cambridge Institute Bioinformatics Core's course on 'Avoiding Data Disasters'

Your Trainers today are:
Mark Fernandes (mark.Fernandes@cruk.cam.ac.uk)
Jing Su
Anne Pajon

We will be using an electronic whiteboard (Etherpad)
Please log onto your computer and access:
https://public.etherpad-mozilla.org/p/2018-2-23-cruk-ci-add

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Welcome to the School of Clinical Medicine's E-Learning suite

- No Eating or Drinking in the suite! We have booked an area in the adjoining corridor with refreshments for a Tea break.
- We are not aware of any planned Fire-drills so if the alarm sounds we will treat it as a real fire and evacuate.
- Ladies and Gentlemans toilets are situated on the level below us (Descend via spiral staircase or lift)
- If you are planning to use your own laptop then you will need to install the free package 'OpenRefine' from http://openrefine.org .
- If you are having difficulties please stick the Red/pink post-it on the back of your display and one of us will come to help you ASAP

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Timetable(ish)

12:30 – 13:10 Introduction & Data formatting* (MF)
13:10 - 14:40 OpenRefine practical (Live coding) (MF+JS+AP)
14:40 – 15:00 Break
15:00– 15:40 File management* (JS)
15:40 – 16:20 Backup & Sharing* (AP)
16:20 - 16:30 Wrap-up & close
* includes 5 mins to write on Etherpad & 5 min look at input

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Data Formatting Issues
Avoiding data disasters –
Best practices in Research Data Management for the Biological Sciences
Feb 23rd 2018

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Reproducible Research

•At some point in the future, someone, somewhere, might want to repeat your analysis for themselves or re-use your data.

- which will most likely be ***you!***

•Assuming that you'll be able to remember all the steps involved is dangerous, so making sure that everything is well-documented is key.

•The documentation involves not only the methods used, but the files used as input and any transformations performed on them.

Five selfish reasons

- Florian Markowetz has a great talk on why we should work reproducibly
- There is a Genome Biology paper that you should read.

Genome Biology

HOME ABOUT ARTICLES SUBMISSION GUIDELINES

COMMENT | OPEN ACCESS

Five selfish reasons to work reproducibly

Florian Markowetz

Genome Biology 2015 16:274 | DOI: 10.1186/s13059-015-0850-7 | © Markowetz. 2015

Published: 8 December 2015

Abstract

And so, my fellow scientists: ask not what you can do for reproducibility; ask what reproducibility can do for you! Here, I present five reasons why working reproducibly pays off in the long run and is in the self-interest of

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

A famous example

•Probably the most (in)famous example of failure to reproduce a study, which actually *put people's lives at risk* and rallied statisticians into action

•Keith Baggerly's lecture on the scandal is a ***must-see***.
https://www.youtube.com/watch?v=7gYIs7uYbMo

• If that wasn't enough to give you sleepless nights –
Visit http://retractionwatch.com

Are spreadsheets programs like Excel evil?

•....Not necessarily.

•Often much more convenient to eye-ball a spreadsheet and get an overall impression of your data.

•But they have *limitations* making them not ideal for large-scale analyses.

•Doing things by-hand only invites you to make copy-and-paste errors etc.

•R cannot read all files as if by magic

Helpful Data Validation features in Excel

- 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

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Less helpful features in Excel

- When identifiers are long integers
  - 1000000 = 1e06
  - Issue with Illumina microaray chip IDs
- Excel can convert gene names to dates
  - SEPT2 (Septin 2) → '2-Sep'
  - MARCH1 (Membrane-Associated Ring Finger (C3HC4) 1, E3 Ubiquitin Protein Ligase) → '1-Mar'

Data Handling rules:

In TV's NCIS, Special Agent Leroy 'Jethro' Gibbs has a set of rules. Failure to observe Gibbs Rules results in a 'Gibbs Slap'

We're not that cruel but here are some rules to follow when dealing with your data.

Image credit:
https://saisoto.deviantart.com/art/Gibbs-Slap-173459786

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Rule 1 -*Never work directly on the raw data*

http://www.inquisitr.com/309687/jesus-painting-restoration-goes-wrong-well-intentioned-old-lady-destroys-100-year-old-fresco/

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Rule 1 -*Never work directly on the raw data*

- Hard to reverse all the manual steps performed and invites errors

- Store the original data somewhere **safe**
    - see later on today

Example 1 – How many inconsistencies can you spot?

Patient ID	Sex	Date of Diagnosis	Tumour Size
1	M	01-01-2013	3.1
2	f	04-18-1998	1.5
3	Male	1st of April 2004	105
4	Female	NA	67
5	F	2010/03/12	4.2
6	F		3.6
7	M	1994-11-05T08:15:30-05:00	232

Rule 2 - Maintain consistency

How many ways can you say “female”?

18-day pregnant females	female (lactating)	individual female	worker caste (female)
2 yr old female	female (pregnant)	lgb*cc females	sex: female
400 yr. old female	female (outbred)	mare	female, other
adult female	female parent	female (worker)	female child
asexual female	female plant	monosex female	femal
castrate female	female with eggs	ovigerous female	3 female
cf.female	female worker	oviparous sexual females	female (phenotype)
cystocarpic female	female, 6-8 weeks old	worker bee	female mice
dikaryon	female, virgin	female enriched	female, spayed
dioecious female	female, worker	pseudohermaprhoditic female	femlale
diploid female	female(gynoecious)	remale	metafemale
f	femele	semi-engorged female	sterile female
famale	female, pooled	sexual oviparous female	normal female
femail	femalen	sterile female worker	sf
female	females	strictly female	vitellogenic replete female
female - worker	females only	tetraploid female	worker
female (alate sexual)	gynoecious	thelytoky	hexaploid female
female (calf)	healthy female	female (gynoecious)	female (f-o)
hen	probably female (based on morphology)		

female (note: this sample was originally provided as a \"male\" sample to us and therefore labeled this way in the brawand et al. paper and original geo submission; however, detailed data analyses carried out in the meantime clearly show that this sample stems from a female individual)",

Courtesy of N. Silvester, European Nucleotide Archive, EMBL-EBI

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Example 1 – applying Rule 2

- Consistency: F, female, f, fem, 2, …
- Units
  - cm or mm; days, months or years
- You can introduce inconsistencies without realising it
  - blank spaces (whitespace) at the end of text
  - "Male " is not the same as "Male"
- Document choices you make about units in a *README* file

Regarding dates
credit: @myusuf3

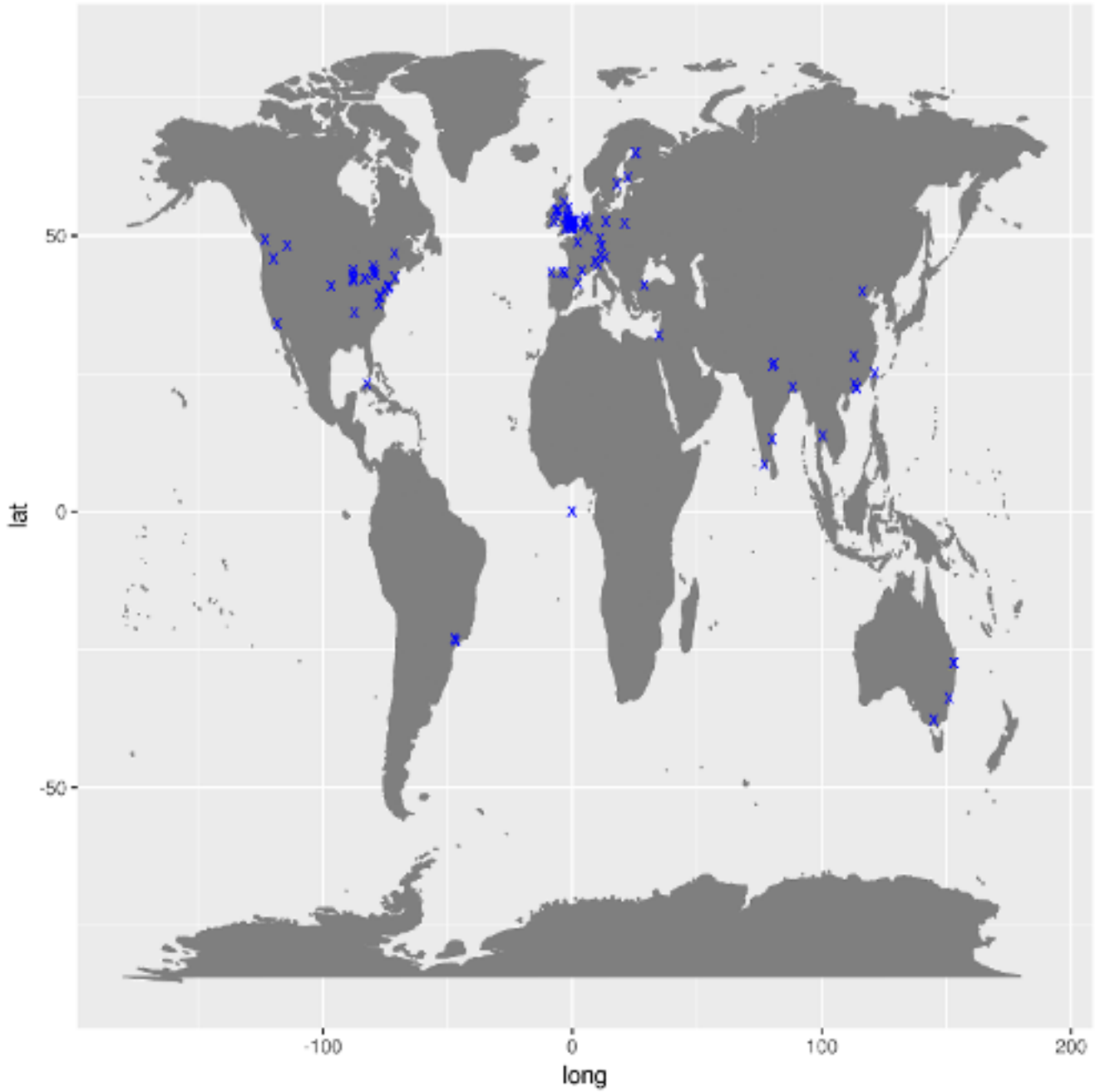
CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Example 1 – corrected using Rule 2

Patient ID	Sex	Date of Diagnosis	Tumour Size
001	M	2013-01-01	3.1
002	F	1998-04-18	1.5
003	M	2004-04-01	1.05
004	F	NA	0.67
005	F	2010-03-12	4.2
006	F	NA	3.6
007	M	1994-11-05	2.32

Rule 3 – Missing values

Figure showing locations of visitors to my Prostate Cancer [data portal]



CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Rule 3 - Don't use 0 to mean missing

- Zero values are data!
  - Sometimes extreme values such as 999 are sometimes used

- NA is Ok, but what if NA is a valid category in your data?
  - R will recognise NA as a missing value and can ignore it in calculations

- Safest to leave the cell *empty*
  - but you need to be careful with blank spaces

Rule 4 - Fill in all the cells

Example 2

Patient ID	Date	Value
1	2015-06-14	213
2		76.5
3	2015-06-18	32
4		120.3
5		109
6	2015-06-20	
7		143

Rule 4 - Fill in all the cells

- It is tempting to make the table look cleaner by not repeating some values
- Fill in all cells!
  - otherwise, problems when sorting
- Empty cell:
  - missing value?
  - value meant to be repeated multiple times?
- Make sure it's clear that the data is missing and not unintentionally left blank

Example 2 Corrected using Rule 4

Patient ID	Date	Value
1	2015-06-14	213
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	
7	2015-06-20	143

Rule 5 - Make it rectangular

- The computer expects a very rigid shape of data with rows and columns

- Each column is a *variable* being examined

- Each row is an observation

- A concept commonly known as *tidy data*

Rule 5 - Make it rectangular

	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

Rule 5 - Make it rectangular

	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

More

- Don’t put too much information in one cell
  - 1 cell = 1 piece of information
- Don’t include units such as "30 g" → "g" in the column name
  - http://unitsofmeasure.org/ucum.html
- Write notes in a separate column or data dictionary or metadata
  - "0 (below threshold)"

More

- Don’t put too much information in one cell
  - 1 cell = 1 piece of information
- Don’t include units such as "30 g" → "g" in the column name
  - http://unitsofmeasure.org/ucum.html
- Write notes in a separate column or data dictionary or metadata
- "0 (below threshold)”

- NO calculations
- NO font colours
- NO highlighting

Computer doesn’t recognize it!

Recapping the ‘Rules’

Rule 1 -Never work directly on the raw data

Rule 2 - Maintain consistency

Rule 3 - Don't use 0 to mean missing

Rule 4 - Fill in all the cells

Rule 5 - Make it rectangular

Write Protection

Mac

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

Write Protection

Windows

- Right click on the file in Windows Explorer
- Properties
- General tab
- Attributes
- Select the box for “read only”

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Practice (makes perfect)

- Look at the file patient-data.txt
  - a simulated, but representative, example of ***bad data***
  - discuss with your neighbours (around 5 minutes)

- The next step is to look at how to clean the data with *Open Refine*

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE