CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

CANCER RESEARCH UK | CAMBRIDGE CENTRE

Data Handling skills

An Introductory session

Introducing James, Mark & Florian

- Dr James Brenton, Brenton group CRUK CI

- Florian Markowetz, Markowetz group CRUK CI

- Mark Fernandes, Bioinformatics Training Developer, CRUK CI

Audience Time!

Enough about us, your opportunity to tell everyone who you are & what you'll be working on and what kinds of data you'll be handling.

Don't be shy!

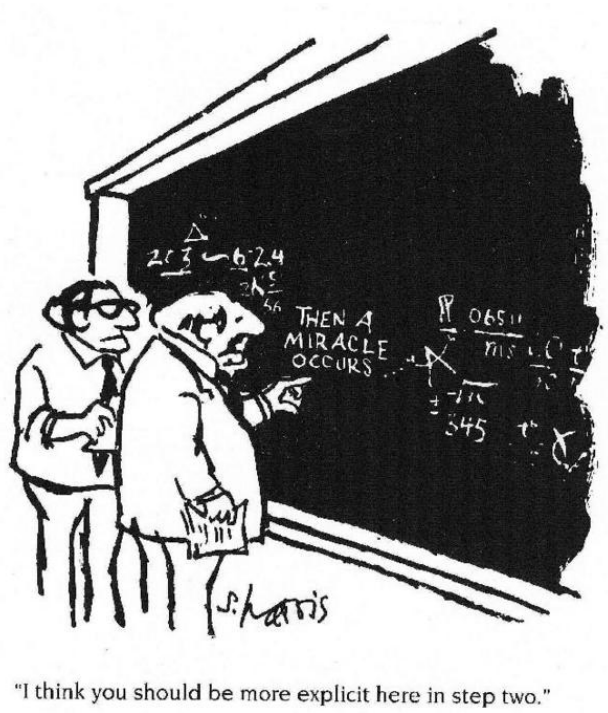
CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

CANCER RESEARCH UK | CAMBRIDGE CENTRE

Flipchart exercise

For your data, what issues/challenges do you anticipate?
Any ideas on how you will meet them?

(Think - Size, confidentiality (GDPR/NHS), provenance, sharing, archive, database creation, file formats…)



"I think you should be more explicit here in step two."

- Goals
  - Reproducibility
  - Robustness
  - (FAIR compliance)
  - Not ending up on Retraction Watch…

- Why work in a reproducible manner?
  – listen to **Florian**!

- Who cares about this? (Funders, University, Journals…)

FAIR Principles | Compliance

Findability
Resource and its metadata are easy to find by both, humans and computer systems. Basic machine readable descriptive metadata allows the discovery of interesting data sets and services.

- **F1.** Resource is uploaded to a public repository.
- **F2.** Metadata are assigned a globally unique and persistent identifier.

Accessibility
Resource and metadata are stored for the long term such that they can be easily accessed and downloaded or locally used by humans and ideally also machines using standard communication protocols.

- **A1.** Resource is accessible for download or manipulation by humans and is ideally also machine readable.
- **A2.** Publications and data repositories have contingency plans to assure that metadata remain accessible, even when the resource or the repository are no longer available.

Interoperability
Metadata should be ready to be exchanged, interpreted and combined in a (semi)automated way with other data sets by humans as well as computer systems.

- **I1.** Resource is uploaded to a repository that is interoperable with other platforms.
- **I2.** Repository meta- data schema maps to or implements the CG Core metadata schema.
- **I3.** Metadata use standard vocabularies and/or ontologies.

Reusability
Data and metadata are sufficiently well-described to allow data to be reused in future research, allowing for integration with other compatible data sources. Proper citation must be facilitated, and the conditions under which the data can be used should be clear to machines

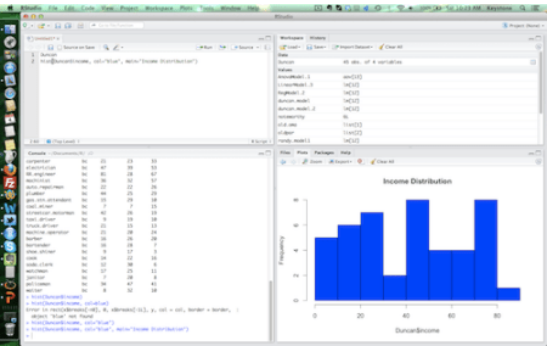
- **R1.** Metadata are released with a clear and accessible usage license.
- **R2.** Metadata about data and datasets are richly described with a plurality of accurate and relevant attributes.

https://retractionwatch.com/2018/11/02/former-university-of-maryland-cancer-researcher-up-to-21-retractions/

Checking data examples – small things, big dividends

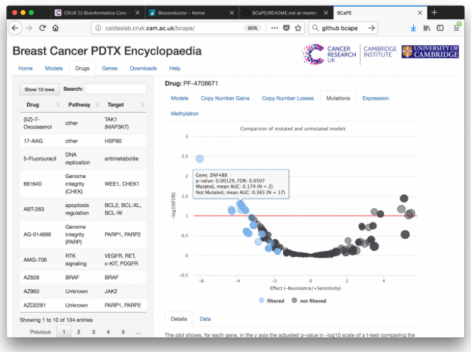
- Manually entered data
  - if possible get double key-entry and use software to compare & find errors
  - Give good documentation on how to enter data & try to anticipate 'special' cases (If you value your data/sanity do NOT let people invent their own codes.).
  - Use Open formats such as CSV or TSV text
  - If using software like Excel use validation option on cells to limit keyable values/types
  - Use tools like OpenRefine to check data before it goes near your analysis workflow
- Replicate checking
  - Do quick plots to perform a quick 'sanity check' of data
    Example of QPCR replicates plots - https://rpubs.com/csaveanu/ggplot_replicates
    - Do lines have a trend – if so, examine any 'odd' outliers

- Tools
  - OpenRefine (http://openrefine.org) demo/activity – fixing supplied data – **Audience**: Can you spot the problems?

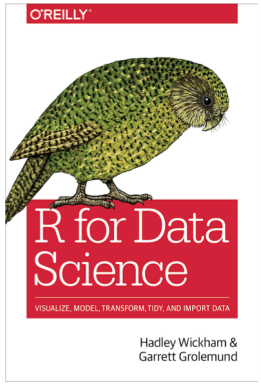


  - R & Rstudio vs. e.g. Excel
    Excel data mangle-ment (Gene ID, Study ID…)
    Automation
    Transparent documented actions (scripts)
  - R packages – Tidyverse (https://www.tidyverse.org)
    GGplot2 and publication ready graphics

  - Repositories (Figshare, Zenodo, Github,…)
  - Databases – ability to share and combine datasets



  - Shiny Apps – publishing models to the web



https://r4ds.had.co.nz

Resources – Who can you go to? (YANA)

- Colleagues & Supervisors
- University Data Champions
- Cam.ac.uk
- Training courses
- Online fora (Stackexchange, Seqanswers, Biostars…)
- Blogs – R-Bloggers

Cambridge Data Champions

https://www.data.cam.ac.uk/intro-data-champions

Conclusions & final advice

- Never work on the raw data – always use a copy
- Document every operation that you do on your data (Use scripts)
- If you're not sure – Stop, put it down and ask someone for advice!
- If you identify a training need, pursue it. We WILL help you.
- Data Champions!
- Always do sanity checks on your data
- Make friends with the Cores (Esp. Bioinformatics ☺) & your Statisticians
- Good idea to learn R – it's a pre-requisite for many courses like the RNASeq one.

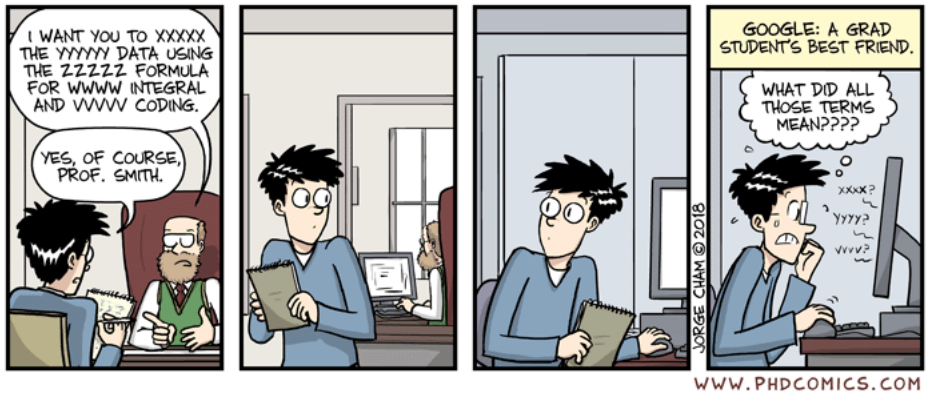
- Finally, Do or do not – there is no try! (Yes I'm a geek).

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

CANCER RESEARCH UK | CAMBRIDGE CENTRE

Thank you!

Any Questions?



http://phdcomics.com

Florians' 5 reasons to work reproducibly lecture:
https://www.youtube.com/watch?v=Is15CMVPHas

This Keith Baggerly lecture is also a must-see.
https://www.youtube.com/watch?v=7gYIs7uYbMo

'Chomping' off units in data using OpenRefine
http://susanemcgregor.com/removing-unwanted-units-from-data-with-chomp-in-google-refine/

Bad (data) Project video
https://www.youtube.com/watch?v=Fl4L4M8m4d0

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

CANCER RESEARCH UK | CAMBRIDGE CENTRE