Introduction to the module.

Data Science option of BIO00058M Data Analysis.

Emma Rand
University of York, UK

Overview

- Aims and learning objectives of 58M
- Pre-module survey results!
- What is Data Science?
  - definition and process
  - reproducibility
  - a rationale for scripting
- Module overview
  - topic list and rationale
  - approach and assessment
  - relationship between topics and assessment

Aims & Learning Outcomes

The aim of 58M *overall* is to enable you to to develop skills in some specific types of 'data analysis' by providing supported practice in workshops and opportunities to apply them independently in 'projects'. This will help you become independent researchers and highly employable.

At the end of this module the successful student will be able to:

1. Demonstrate the acquisition of skills in experimental design and data analysis, related to the option chosen within the module.
2. Apply the skills learned to address novel bioscience problems.

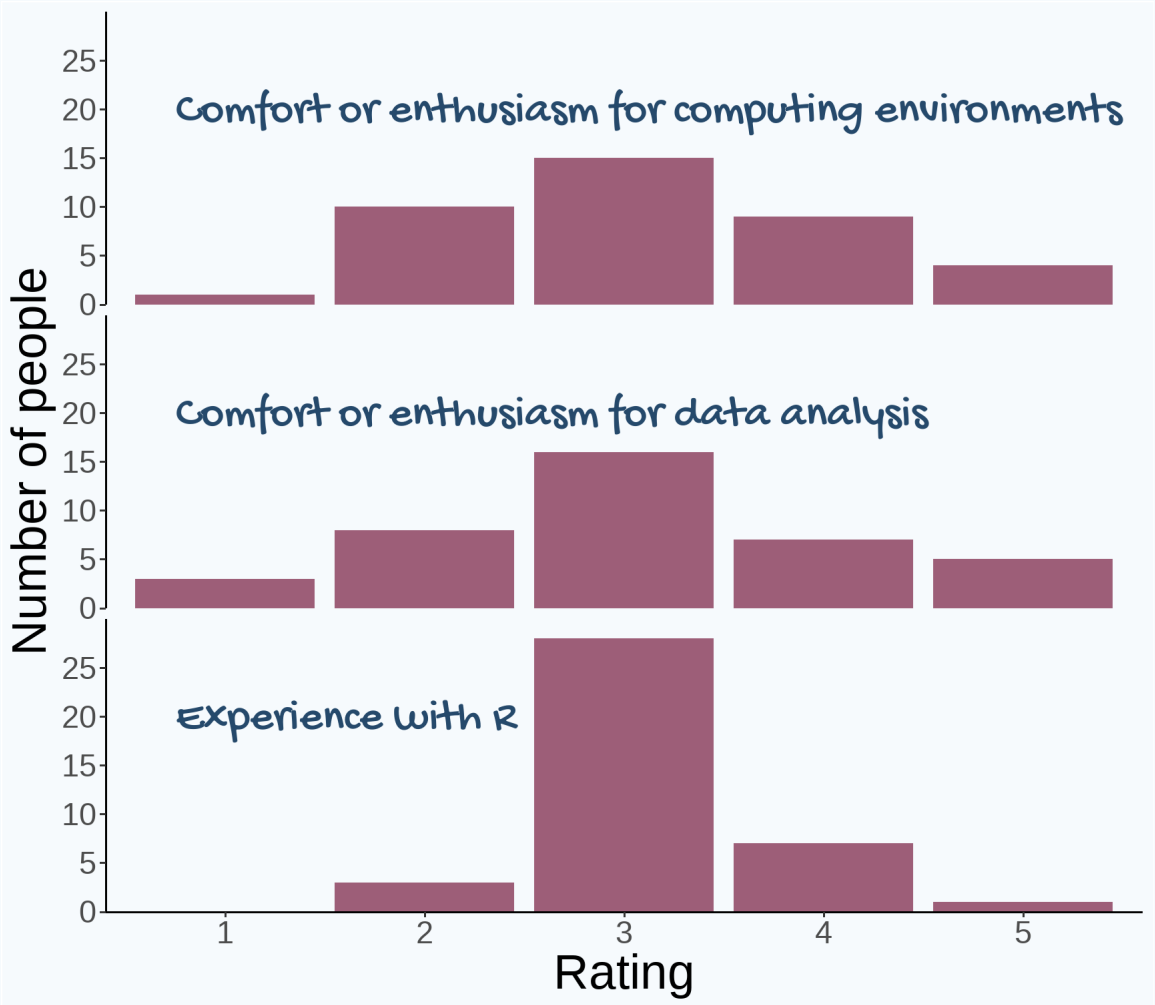
Learning Outcomes of Data Science

For Data Science, the first objective means:

Produce a reproducible data analysis and report. The analysis can emphasise data import, processing, statistical analysis, visualisation or any combination of these.

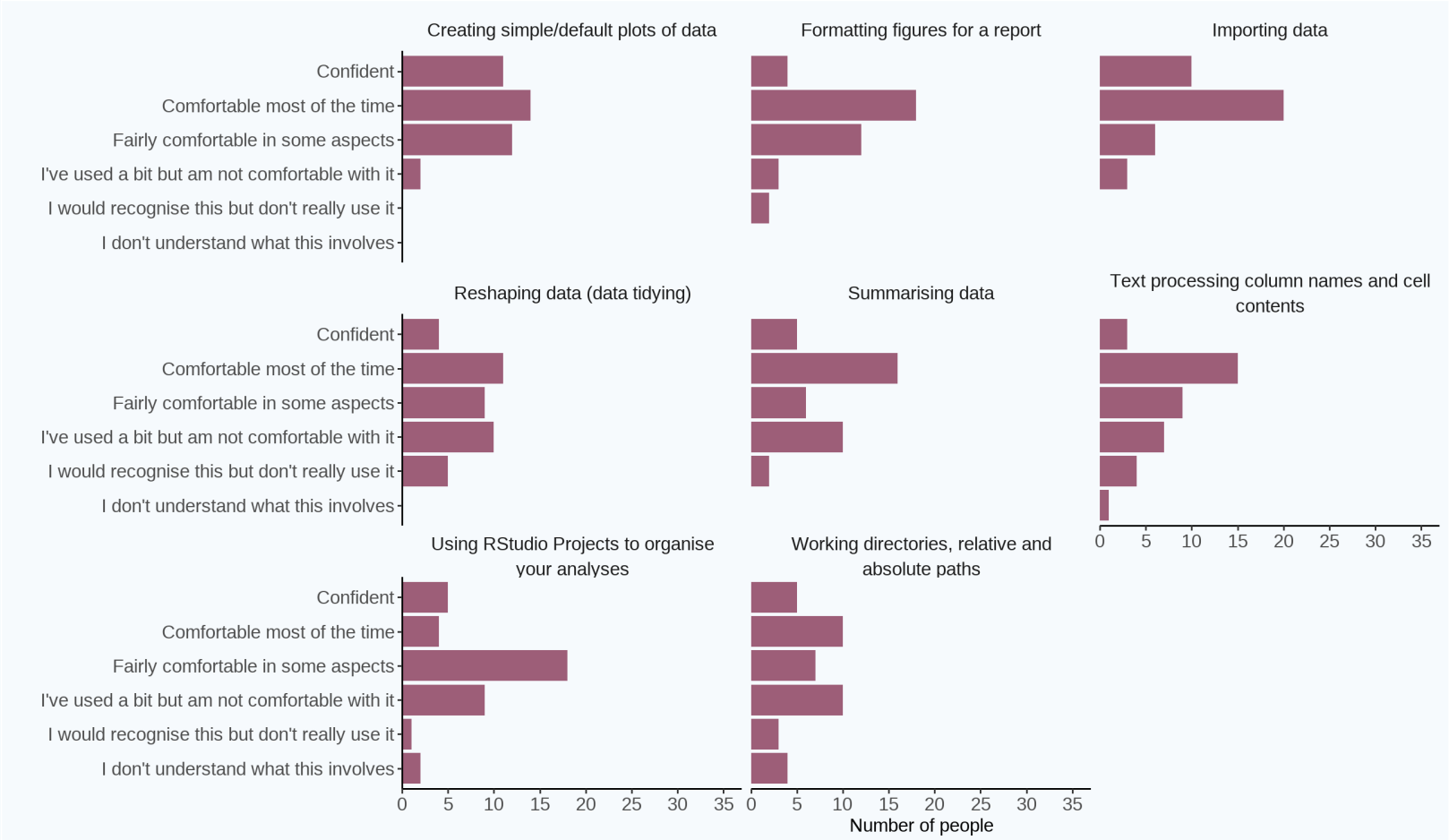
Survey results 1/3

The distribution of ratings you (n = 39) gave in the survey were:



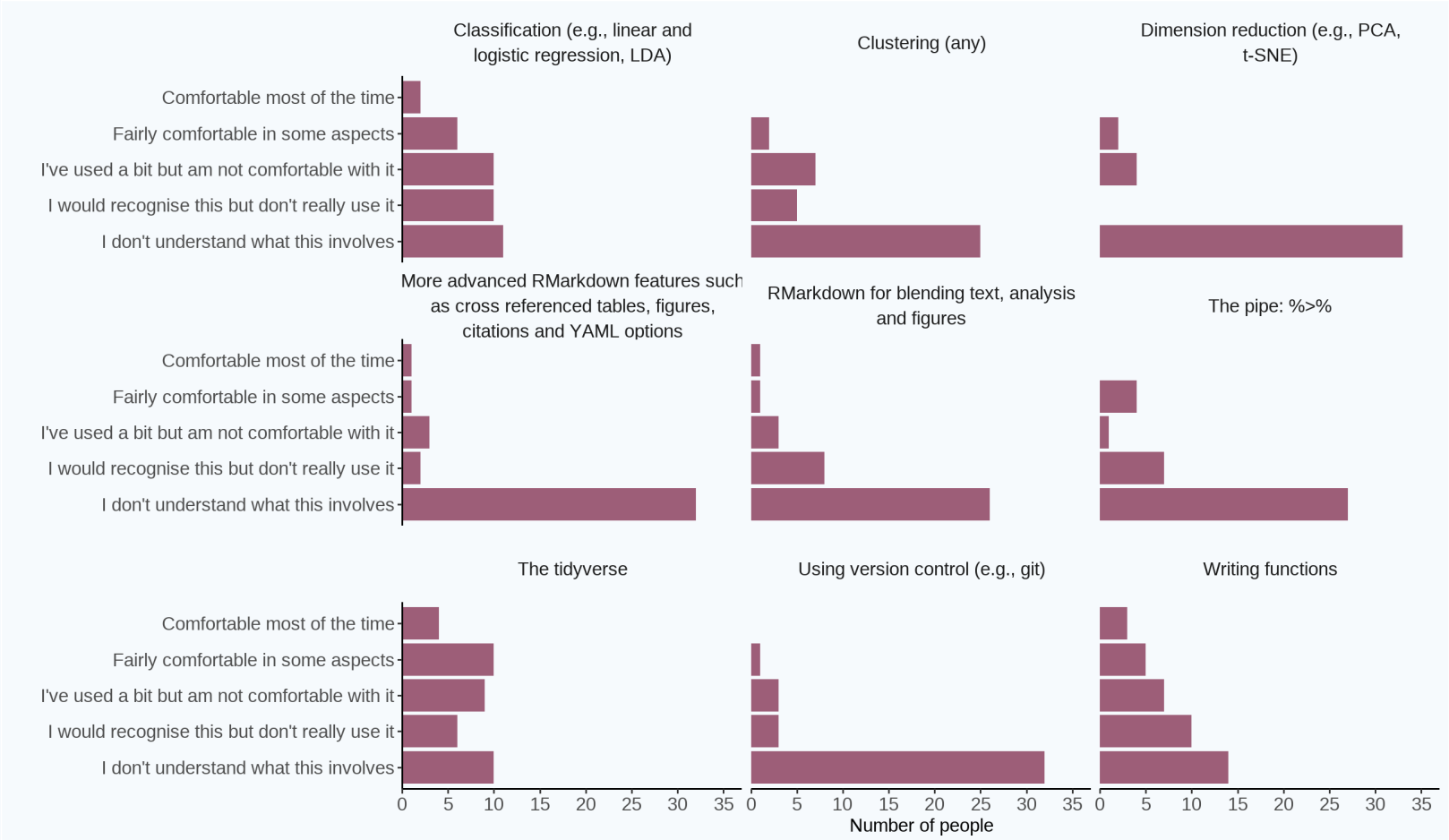
Survey results 2/3

Please rate your level of comfort with...



Survey results 3/3

Please rate your level of comfort with...



What is Data Science?

The development, and application, of reproducible workflows for the simulation, collection, organisation, processing, analysis and presentation of data in order to extract knowledge or insight.

Data science underlies open and reproducible research.

How much of data science is using statistics?

Less than you probably think

~80% of your time on getting data, cleaning data, aggregating data, reshaping data, and exploring data using exploratory data analysis and data visualization.

Reproducibility is key!

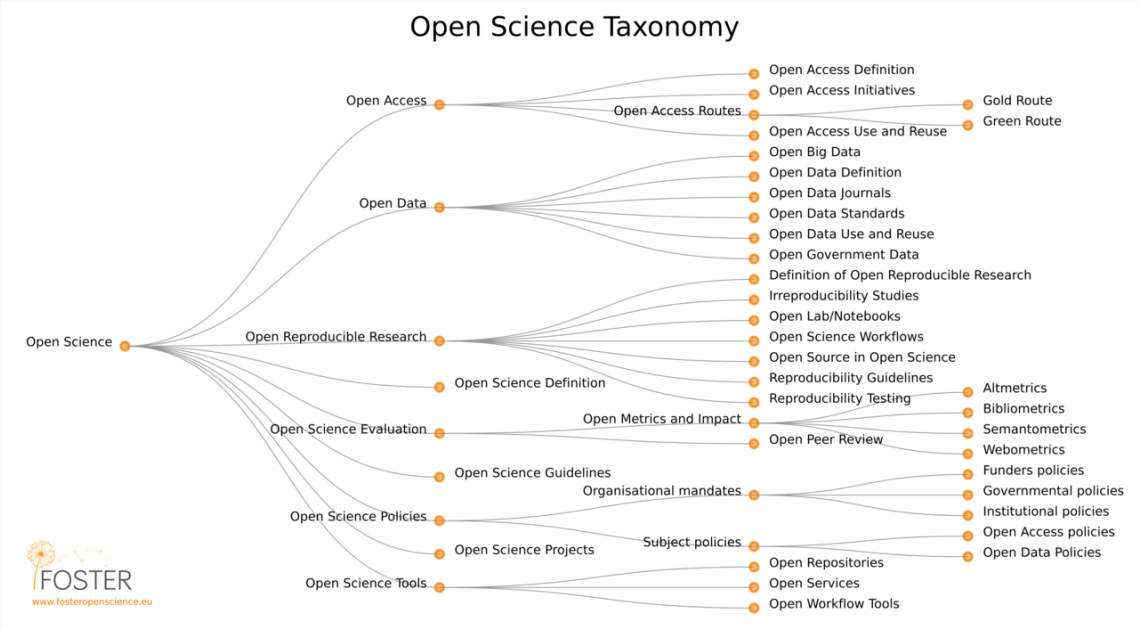
One definition *"... obtaining consistent results using the same input data; computational steps, methods, and code; and conditions of analysis. This definition is synonymous with "computational reproducibility"*. (National Academies of Sciences, Engineering Medicine, et al., 2019)

Also see National Science Foundation (Bollen Cacioppo, et al., 2015)

Who cares?

- Many high profile cases of work which did not reproduce e.g. Anil Potti unravelled by Baggerly and Coombes (2009)
- Five selfish reasons to work reproducibly (Markowetz, 2015b). Alternatively, see the talk
- Will become standard in Science and publishing e.g OECD Global Science Forum Building digital workforce capacity and skills for data-intensive science (OECD Global Science Forum, 2020)

Open Science

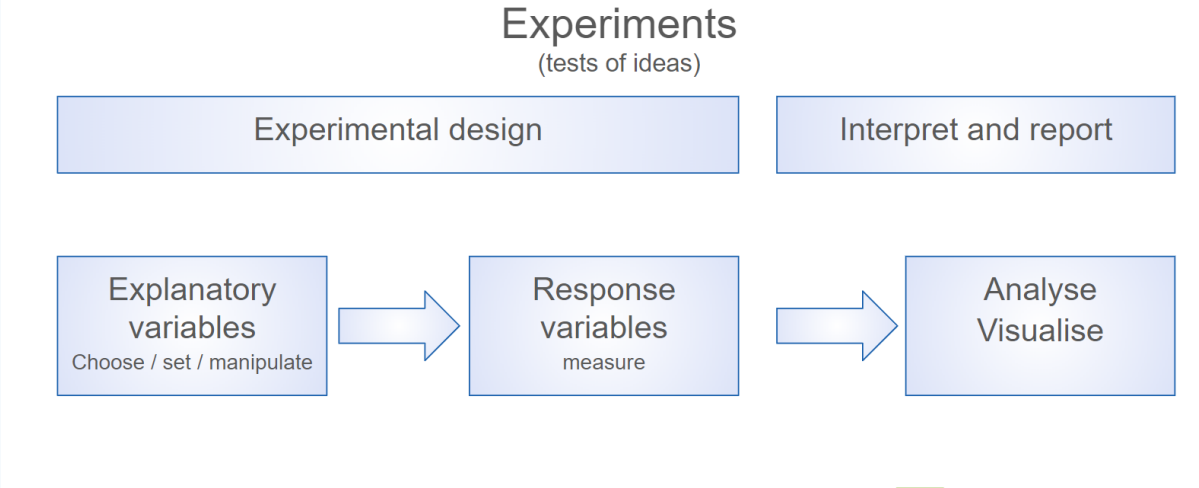


By Petr Knoth and Nancy Pontika - https://en.wikipedia.org/wiki/Open_science#/media/File:Os_taxonomy.png, CC BY 3.0, https://commons.wikimedia.org/w/index.php?curid=61125075

FAIR - Findable, Accessible, Interoperable, Reusable (Wilkinson Dumontier, et al., 2016)

Rationale for scripting analysis

Science is the generation of ideas, designing work to test them and reporting the results.

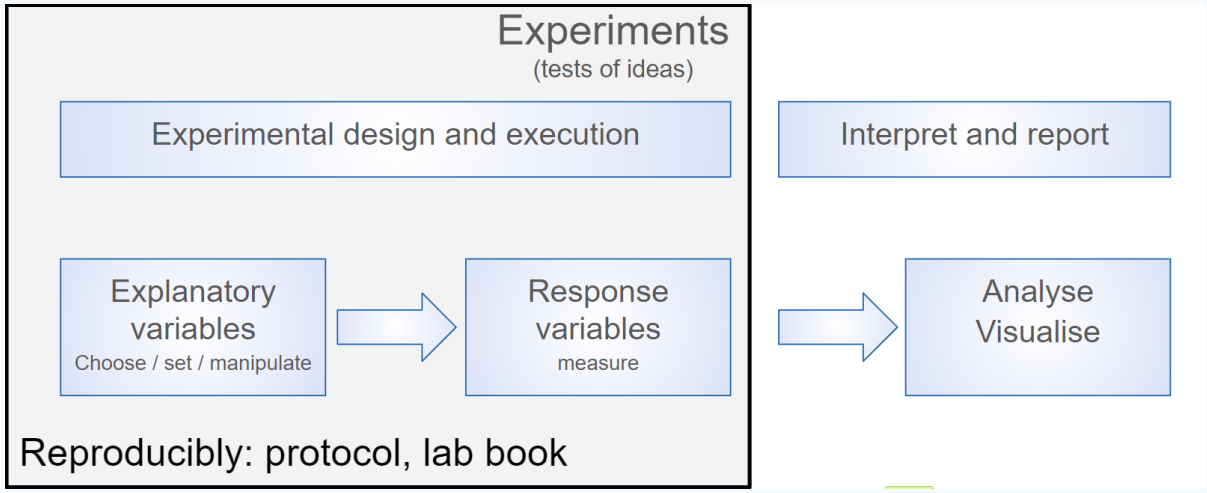


Generating the results

Analysing and reporting them

Rationale for scripting analysis

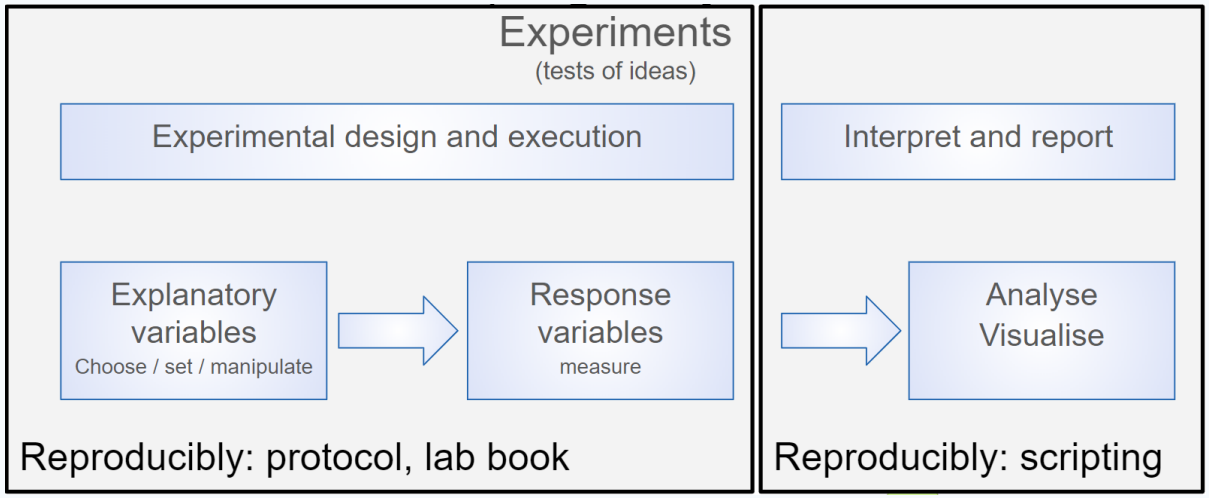
We ensure reproducibility of laboratory and field work by planning and recording in lab books and using standard protocols.



Even so replicating results can be hard.

Rationale for scripting analysis

We ensure reproducibility of laboratory and field work by planning and recording in lab books and using standard protocols.

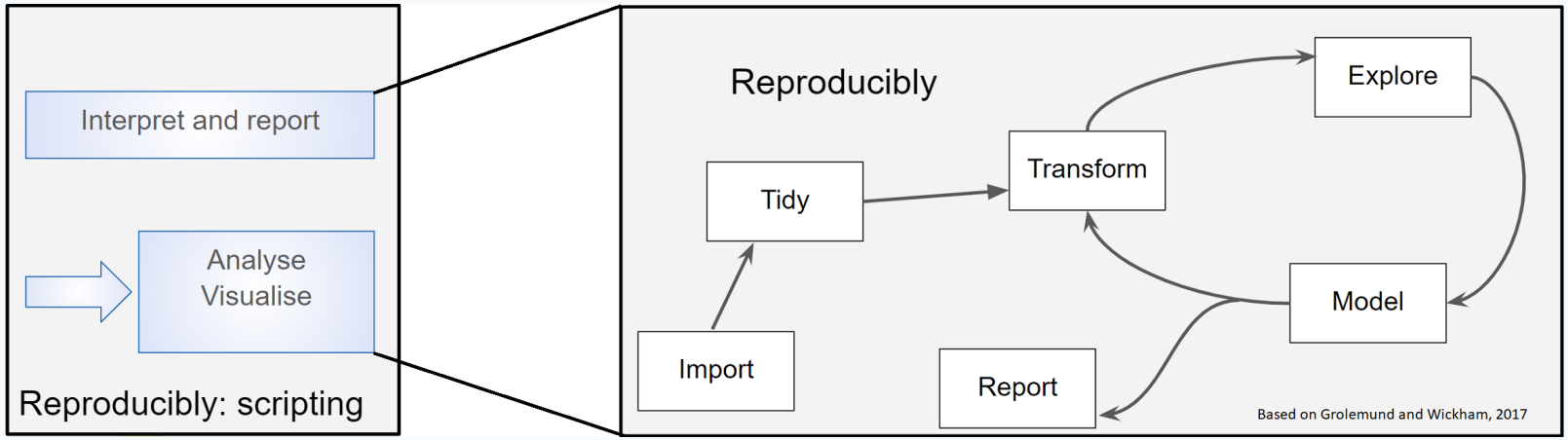


Even so replicating results can be hard.

Workflows for computational projects and the data analysis and reporting of other work can, and should, be 100% reproducible! Scripting is the way to achieve this.

Rationale for scripting analysis

That reproducibility applies to all aspects of the data workflow.



From importing or collecting the data, processing it for analysis, building statistical models and communicating the methods and results.

These are usually iterative and that process of iteration (the development of the analysis) should also be captured.

Why R?

Open source and free .....But so is Python

R has reputation for catering to users who do not see themselves as programmers, and allowing them to slide gradually into programming.

It was also designed for data analysis and graphics - it is a 'domain specific' programming language - which means it is usually easier to achieve those tasks in R than a general purpose programming language.

Why R?

The R community is one of R's greatest assets, being vibrant, inclusive and supportive of users at all levels.

- #rstats on twitter is very active
- RForwards the widening participation task force [1]
- RLadies gender diversity promotion
- Hey! You there! You are welcome here

Artwork by @allison_horst "welcome to rstats on twitter"

1. I am member of the Core Team for Forwards

Why R?

R Markdown is sometimes called R's "killer feature". It turns your analyses into fully reproducible high quality documents, reports, presentations and dashboards.

Artwork by @allison_horst "Be an Rmarkdown knitting wizard."

Module overview

Chosen topics are: foundational, follow stages 1 and 2 well, are widely applicable (in this module and beyond) and transferable conceptually:

- Using RStudio projects and an emphasis on good practice in code and project documentation and organisation including version control and collaboration.
- More advanced data tidying.
- An emphasis on reproducibility and reproducible reporting using R Markdown.
- Some machine learning concepts and methods that are very commonly applied independent of the data domain.

You will also have the time and opportunity to independently develop skills particular to your interests and the assessment undertaken with support.

Week plan

- Week 2: Preparation 1 - Update your R and RStudio, revise previously taught material.
- Week 3: Preparation 2 - Introduction to the module. Installing git and getting a GitHub account.
- Week 4: Topic 1 - Project organisation. Tools for version control and collaborating. *
- Week 5: Topic 2 - Tidying data and the tidyverse.*
- Week 6: Topic 3 - Reproducibility and an introduction to R Markdown.*
- Week 7: Topic 4 - Advanced R Markdown.*
- Week 8: Topic 5 - An introduction to Machine Learning: Overview and Unsupervised methods.*
- Week 9: Topic 6 - An introduction to Machine Learning: Supervised Methods.*
- Week 10: Project consultation. *
- Spring Weeks 1 - 5: Project work. Drop-ins TBC

* week includes a timetabled online session

Approach and Assessment

I wanted you to

- learn essential skills for reproducible and open research but otherwise..
- be able to work on problems you are interested in and/or that are related to project or module work
- have time to develop the skills for that, with support
- be assessed on what you can do (not what you can't do)

Thus there is choice and flexibility in the assessment. You can chose your own data set and problem. Two highly scoring submissions could look completely different.

The options

1. Reproducible analysis related to your project/module work
    - Analysis of existing or simulated data including images
    - Conversion of existing lab tools (eg excel files) to reproducible pipelines
    - Analysis of literature (text analysis)
2. Reproducible analysis of previous work undertaken unreproducibly
    - 58I Bioscience Techniques - almost all of the analyses undertaken (use of excel, Summit, ImageJ etc) can be coded reproducibly.
3. Reproducible analysis of a provided project - because not everyone enjoys choosing their own

Topics and assessment

- Week 4: Topic 1 - Project organisation.

  Tools for version control and collaborating.
- Week 5: Topic 2 - Tidying data and the tidyverse.
- Week 6: Topic 3 - Reproducibility and an introduction to R Markdown.
- Week 7: Topic 4 - Advanced R Markdown.
- Week 8: Topic 5 - An introduction to Machine Learning: Overview and Unsupervised methods.
- Week 9: Topic 6 - An introduction to Machine Learning: Supervised Methods.

In your assessment you **must** use RStudio projects and rmarkdown, organise your analyses for reproducibility and follow good practice.

Topics and assessment

- Week 4: Topic 1 - Project organisation.

  Tools for version control and collaborating.
- Week 5: Topic 2 - Tidying data and the tidyverse.
- Week 6: Topic 3 - Reproducibility and an introduction to R Markdown.
- Week 7: Topic 4 - Advanced R Markdown.
- Week 8: Topic 5 - An introduction to Machine Learning: Overview and Unsupervised methods.
- Week 9: Topic 6 - An introduction to Machine Learning: Supervised Methods.

The extent of data tidying and processing and machine learning methods will vary depending on your project.

Version control is not required - it's taught only to help you collaborate online.

Topics and assessment

- Week 4: Topic 1 - Project organisation.

  Tools for version control and collaborating.
- Week 5: Topic 2 - Tidying data and the tidyverse.
- Week 6: Topic 3 - Reproducibility and an introduction to R Markdown.
- Week 7: Topic 4 - Advanced R Markdown.
- Week 8: Topic 5 - An introduction to Machine Learning: Overview and Unsupervised methods.
- Week 9: Topic 6 - An introduction to Machine Learning: Supervised Methods.

You can learn something completely different as long as your work is reproducible and uses an RMarkdown format. This includes shiny applications.

Assessment

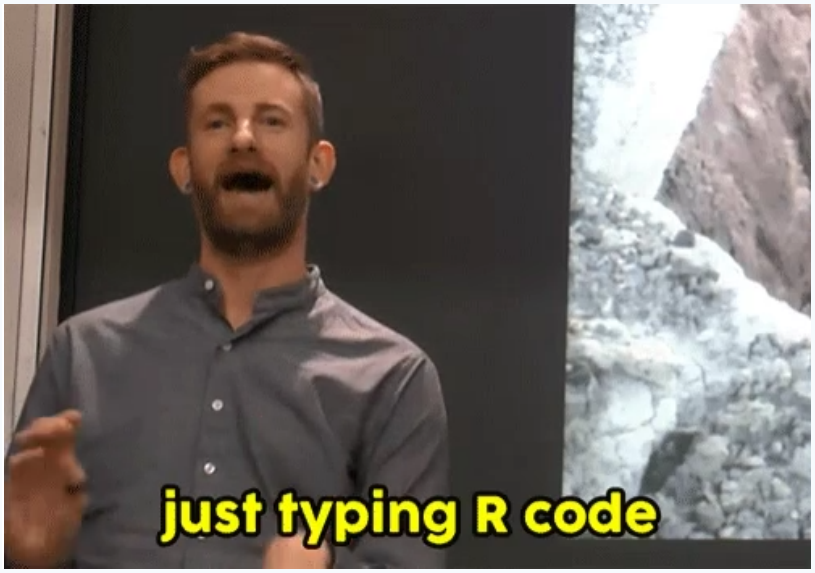
Everyone will have the opportunity for a one-to-one meeting to discuss the remit and scoping of their project during the Autumn term. Book 58M Project one-to-one

Everyone will have the opportunity for a one-to-one meeting for formative feedback on their project during the Spring term.

Questions

You can ask any questions about taught materials or the assessment in the Blackboard Collaborate workshop chat.

You can also ask any questions here



References

Baggerly, K. A. and K. R. Coombes (2009). "DERIVING CHEMOSENSITIVITY FROM CELL LINES: FORENSIC BIOINFORMATICS AND REPRODUCIBLE RESEARCH IN HIGH-THROUGHPUT BIOLOGY". In: *Ann. Appl. Stat.* 3.4, pp. 1309-1334.

Bollen, K., J. T. Cacioppo, et al. (2015). *Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science*. National Science Foundation.

Markowetz, F. (2015b). "Five selfish reasons to work reproducibly". En. In: *Genome Biol.* 16, p. 274.

National Academies of Sciences, Engineering, Medicine, et al. (2019). *Understanding Reproducibility and Replicability*. National Academies Press (US).

References

OECD Global Science Forum (2020). *Building digital workforce capacity and skills for data-intensive science*. OECD.

Wilkinson, M. D, M. Dumontier, et al. (2016). "The FAIR Guiding Principles for scientific data management and stewardship". En. In: *Sci Data* 3, p. 160018.

Xie, Y. (2019). *xaringan: Presentation Ninja*. R package version 0.12. URL: https://CRAN.R-project.org/package=xaringan.

Slides made with with xaringan (Xie, 2019) and xaringanExtra (Aden-Buie, 2020)

Organisations

- The Alan Turing Institute
- Software Sustainability Institute
- UK Reproducibility Network
- FOSTER Plus
- Center for Open Science

Emma Rand
emma.rand@york.ac.uk
Twitter: @er13_r
GitHub: 3mmaRand
blog: https://buzzrbeeline.blog/