Biological Data Science using R

Lecture 1: Introduction

Options: You do one of:

Choice 1	Stage 3	Stage 4	Grand Total
Analysing and using 3D structures	11	5	16
Biological Data Science	22	30	52
Image Analysis	16	20	36
Sequence analysis	23	19	42

Each option is about 15 hours contact time

Assessment criteria are the same

Lecture Overview

- Aims and LO of 58M
- Survey results
- What is data science?
  - Definition, process
  - Reproducibility
  - Rationale for scripting
- BDS overview
  - Topic rationale
  - Session list
  - My objectives
  - Approach, assessment and the LO
  - Relationship between sessions and assessment
- Questions!

Aims and 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.

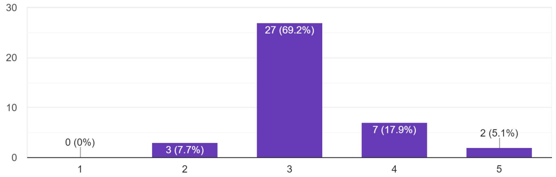
For BDS, 1. means:

1. Devise reproducible strategies to import, tidy, transform, model and report on data in R.

Survey results

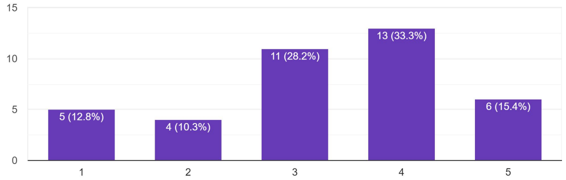
Rate your level of experience with R relative to your peers. Consider 3 to be about average, for someone who has ha...t for data analysis in other modules.

39 responses



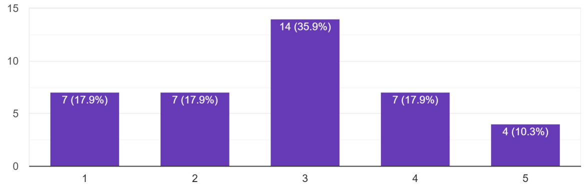
Rate your comfort and enthusiasm for data analysis in general relative to your peers

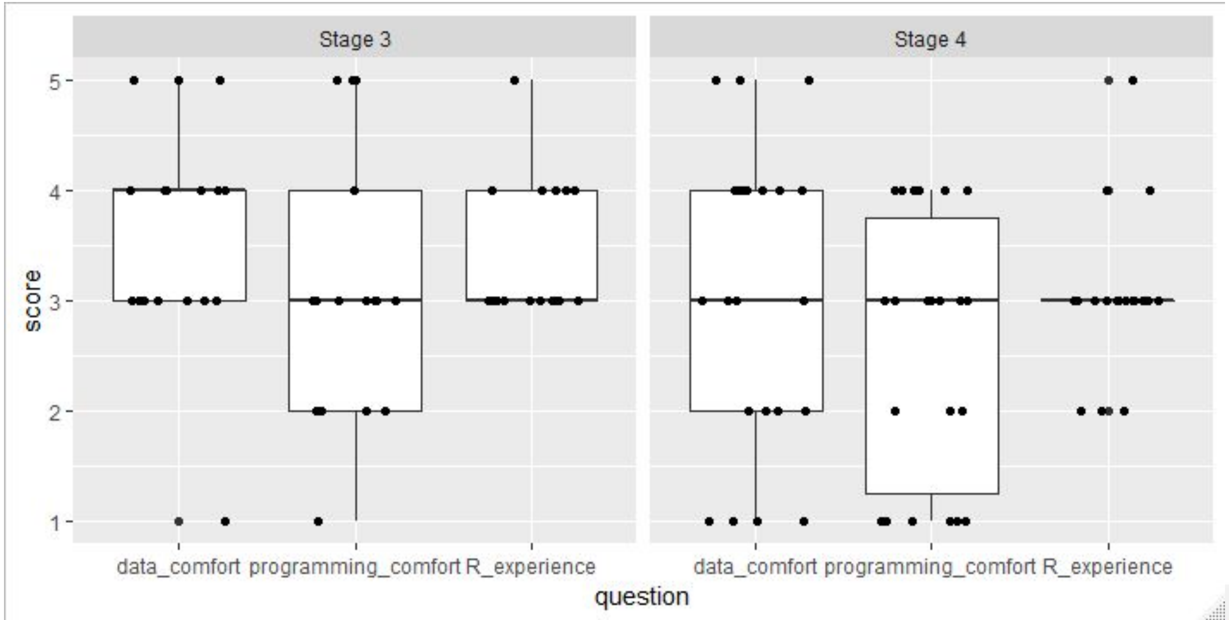
39 responses



Rate your comfort or enthusiasm for programming relative to your peers

39 responses





What is Data Science?

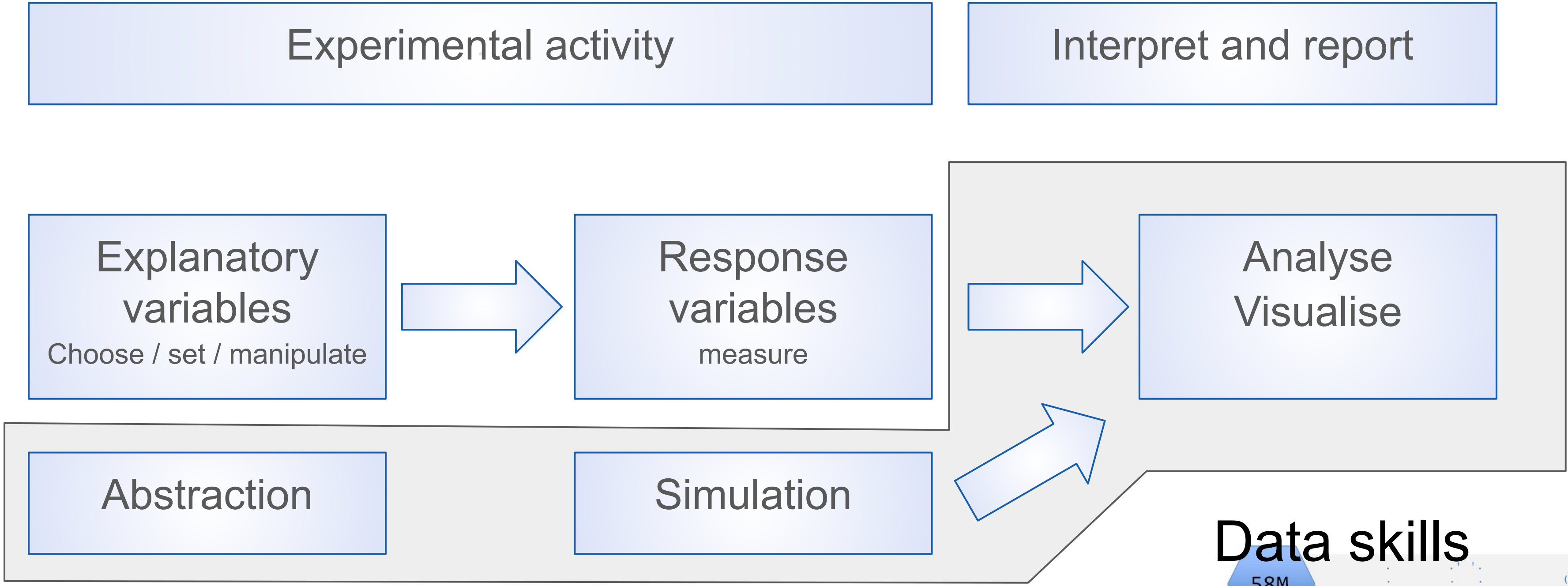
Not the same as numeracy - you don’t have to be good at maths
Not the same as Statistics: includes statistical analysis but also what you have to do before and after.

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

Science

Experiments

(tests of ideas)



Experimental activity

Interpret and report

Explanatory variables

Choose / set / manipulate

Response variables

measure

Analyse

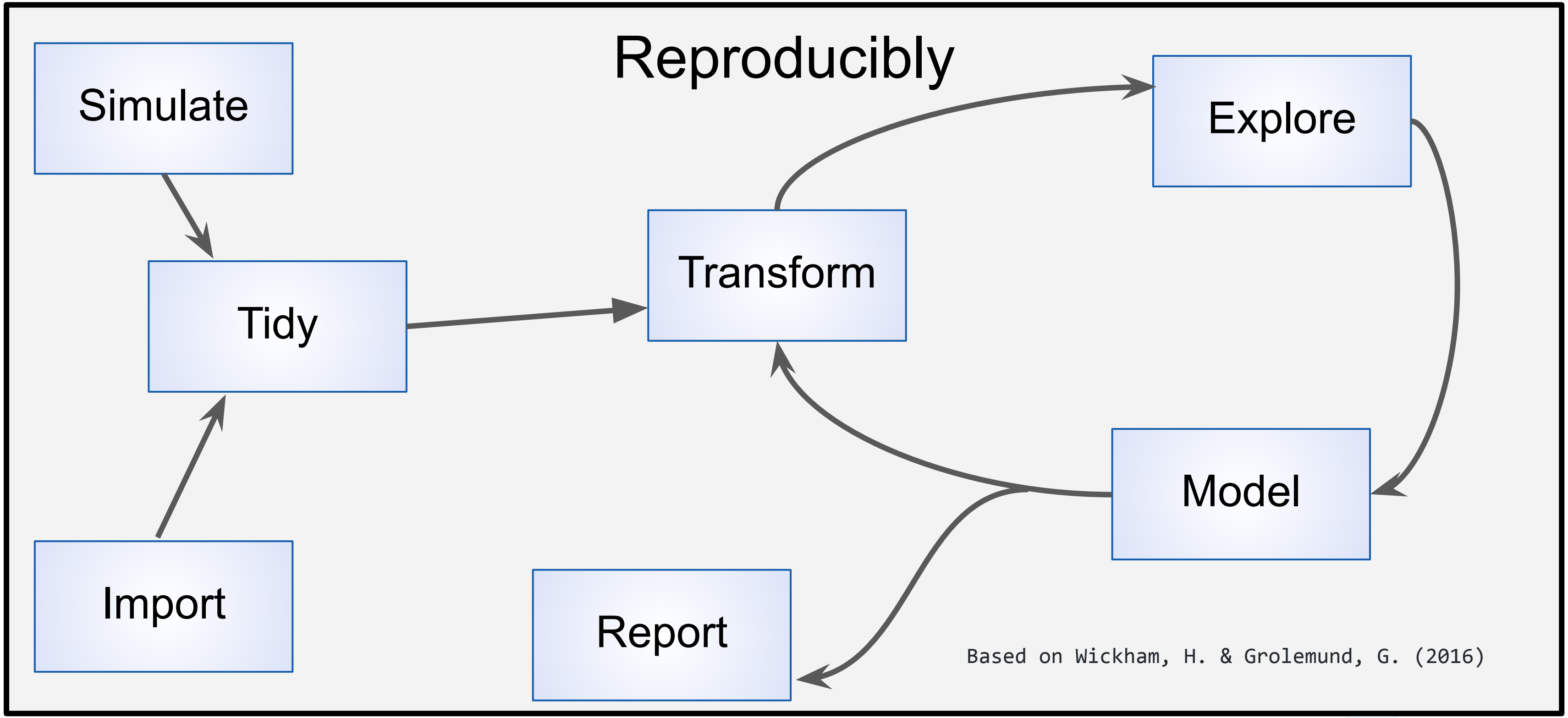
Visualise

Abstraction

Simulation

Data skills

What is data science



Reproducibly

Simulate

Tidy

Import

Transform

Explore

Model

Report

Based on Wickham, H. & Grolemund, G. (2016)

How much of data science is using statistics?

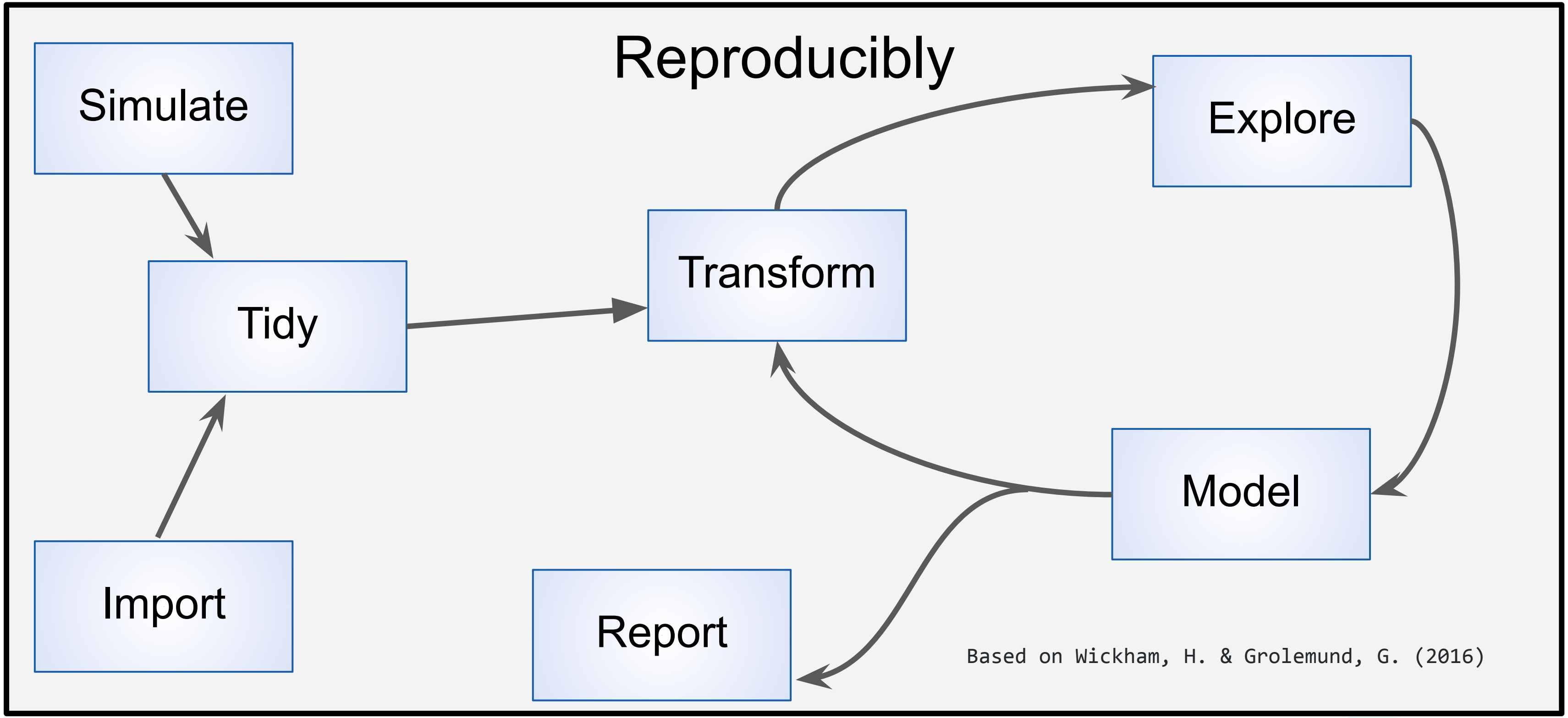
Less than you probably think

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

Data analysis means: getting data, reshaping it, exploring it, and visualizing it as well as modelling

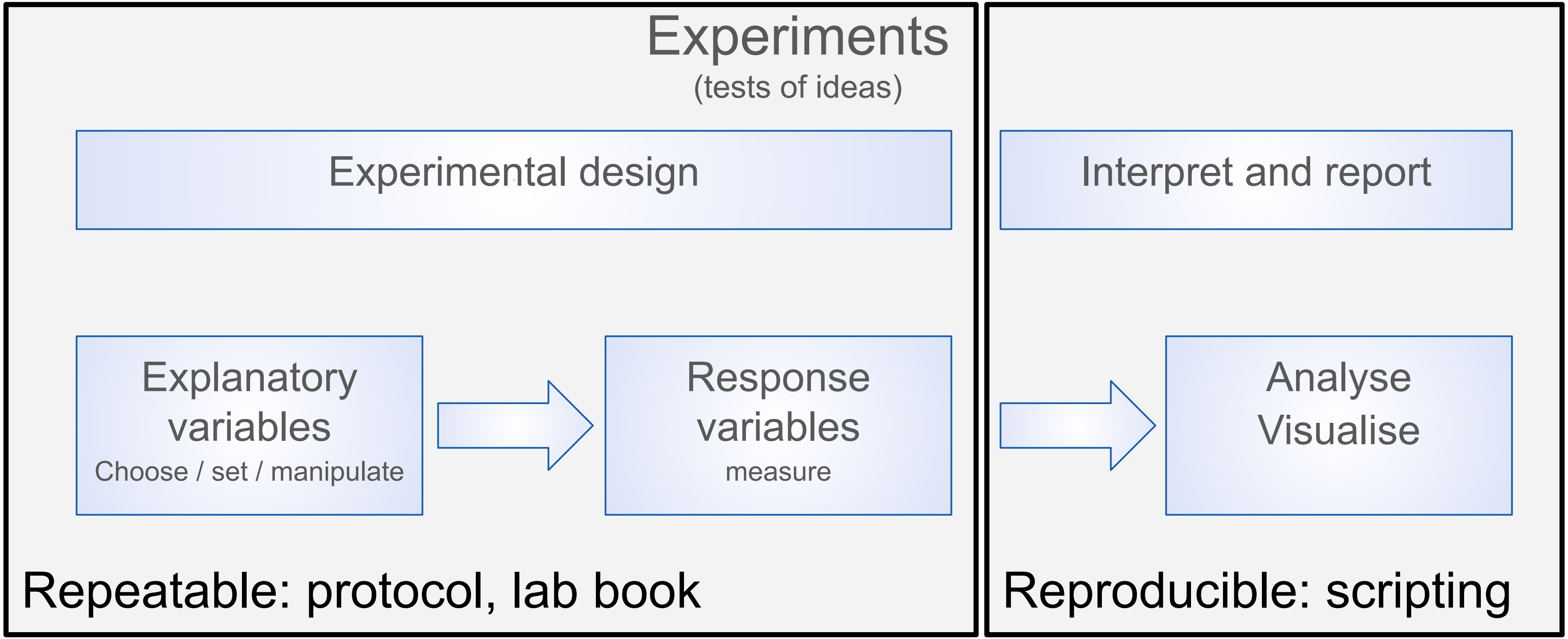
Reproducibility: same data + same analysis = same results

Reproducibility is a key feature



Based on Wickham, H. & Grolemund, G. (2016)

Rationale for scripting



Experiments

(tests of ideas)

Experimental design

Explanatory variables

Choose / set / manipulate

Response variables

measure

Repeatable: protocol, lab book

Interpret and report

Analyse

Visualise

Reproducible: scripting

Reproducible, Repeatable, Replicated

Replication: within a study

Repeatable: between studies. Independently, without the use of original data but generally using the same methods.

Reproducible: The original data and original methods reproduce all of the findings of a study.

Methods need to be perfectly described

Patil et al. A statistical definition for reproducibility and replicability

Overview of Module

Impossible to cover everything to you might ever need! Different people will use different methods and tools.
Chosen topics are: foundational, follow stages 1 and 2 well, widely applicable (in this module and beyond), transferable conceptually.
Those topics are

- using RStudio projects and good practice in organisation.
- more advanced data importing and 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 have the time and opportunity to independently develop skills particular to your interests and the assessment undertaken with support.

Sessions (ignore timetable naming!)

Workshop 1: Project Organisation.

Workshop 2: Tidying data and the tidyverse.

Workshop 3: Advanced Data Import.

Workshop 4: Reproducibility and an introduction to R Markdown.

Workshop 5: Advanced R Markdown.

Workshop 6: An introduction to Machine Learning.

Workshop 7: Project work

Drop-ins: 3 x 2hrs (unfortunately 1600 - 1800 Friday)

My objectives!

Create a learning environment characterised by

- A focus on progress and improvement
- Enjoyment and satisfaction
- Interaction and exchange of ideas
- Initiative and independence
- Supported problem solving

Cater equally well to stage 3 and stage 4 students.

Assessment, learning objectives and approach

I didn’t want

- one size fits all
- artificial/meaningless jumping through hoops
- fear of failure and judgement to interfere

I did want you to

- be able to work on problems you are interested in
- be able to develop the skills needed for that
- have more supported unstructured time
- be assessed on what you can do (not what you can’t do)

Core skills taught with examples in 6 workshops which should also have time to for you to apply to your own work.
Support to learning how to create a reproducible analysis related to your project, a past ‘project’ of or provided ‘projects’

Assessment

Choice!

1. Reproducible analysis related to your project
   a. Analysis of existing or simulated data including images
   b. Conversion of existing lab tools (eg excel files) to reproducible pipelines
   c. Analysis of literature
2. Reproducible analysis of previous work undertaken unreproducibly
   a. 58I (32I) Bioscience Techniques option - cell imaging (ImageJ), flow cytometry (Summit), pbd files, excel files
3. Reproducible analysis of a provided project: VLE

The submission is a zip of the whole project - rmd, output, accessory scripts, data. Examples are on the VLE.The Rmd should be well-commented and contain everything needed to recreate, and understand the recreation of, the knitted output.The knitted output should be no more than 2000 words.

Relationship between sessions and assessment

Workshop 1: Project Organisation.

Workshop 2: Tidying data and the tidyverse.

Workshop 3: Advanced Data Import.

Workshop 4: Reproducibility and an introduction to R Markdown.

Workshop 5: Advanced R Markdown.

Workshop 6: An introduction to Machine Learning.

Relationship between sessions and assessment

Workshop 1: Project Organisation.

Workshop 4: Reproducibility and an introduction to R Markdown.

Workshop 5: Advanced R Markdown.

All assessments must implement ideas cover in these sessions

Relationship between sessions and assessment

Workshop 2: Tidying data and the tidyverse.

Workshop 3: Advanced Data Import.

Workshop 6: An introduction to Machine Learning.

You may implement some of these more than others. You need to go beyond what was directly taught in these or some other aspect (e.g., image analysis, ‘omics).

Think about what you want to work on

https://forms.gle/3CTGcwUqB7GPvABU7

Discuss with me to confirm

Confirm by Workshop 2 (11 Oct)?

We have some time to discuss now