

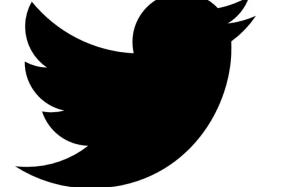
The three R's with R

Research data management
Reproducibility
Reusability

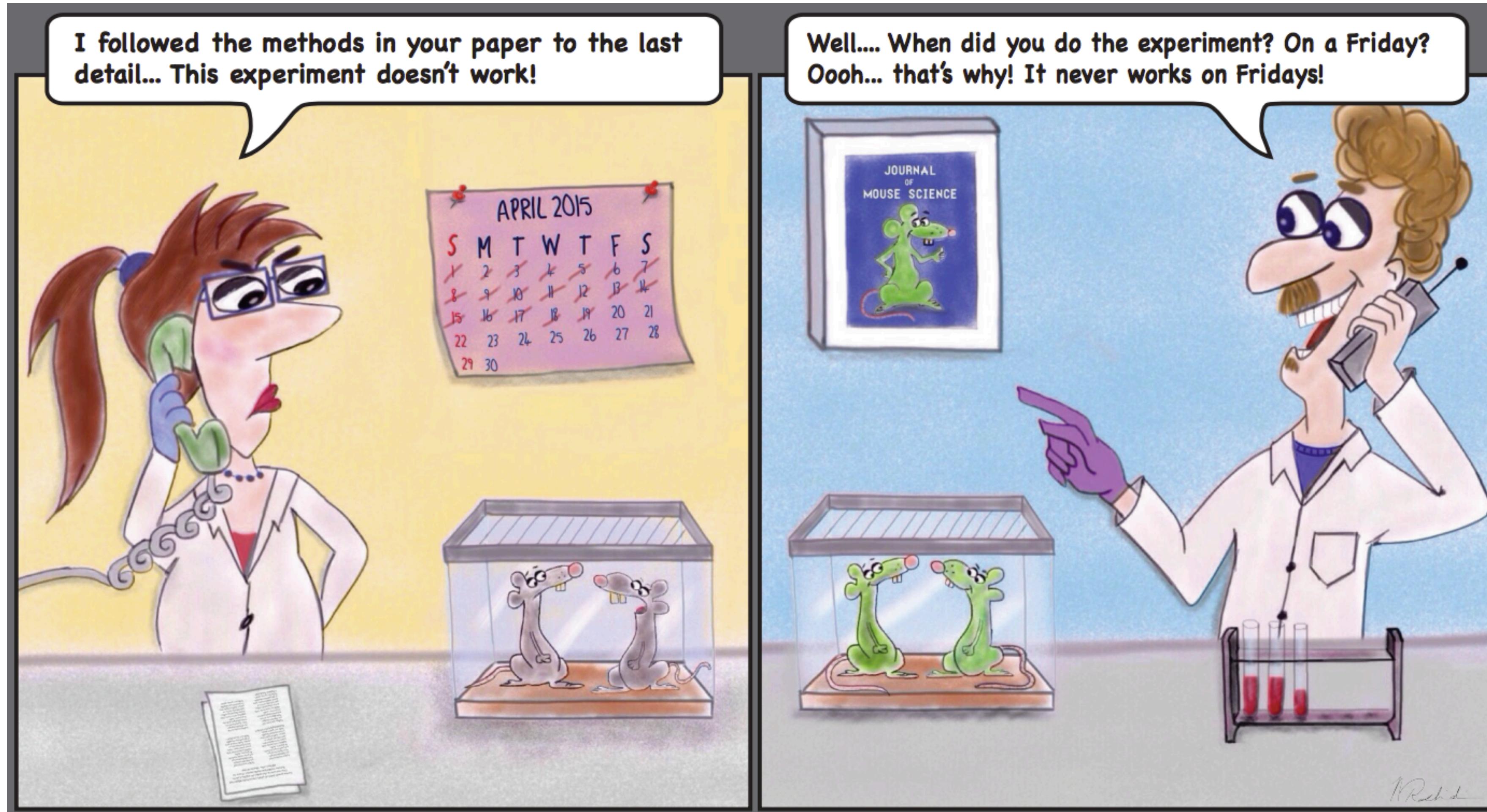


Beautiful data for all the world to see

Dr Jessica Stapley
jessica.stapley@env.ethz.ch

 @jessstapley

Reproducibility is a major principle of the scientific method



Reproducibility versus Replication

Reproducibility

An independent researcher can obtain the same results using the original data and original computer code

Replication (Independent verification)

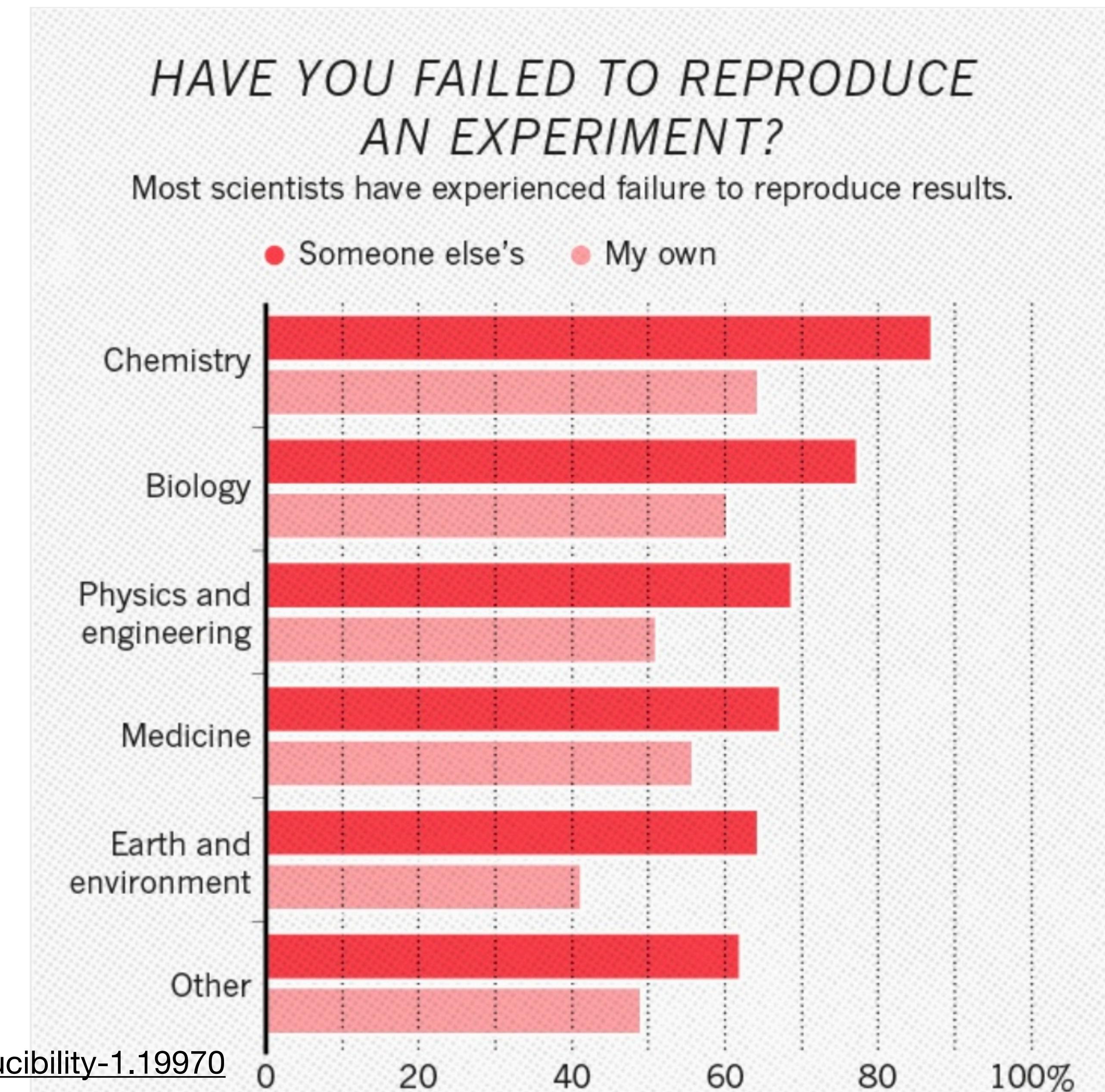
Different researcher/lab conducts an independent study and obtains the same results or comes to the same conclusion as the original study

But how reproducible are most scientific findings?

Not very

Survey of >1500 researchers by Nature

- >70% of researchers have tried and failed to reproduce another lab's experiments
- >50% failed to reproduce their own experiments



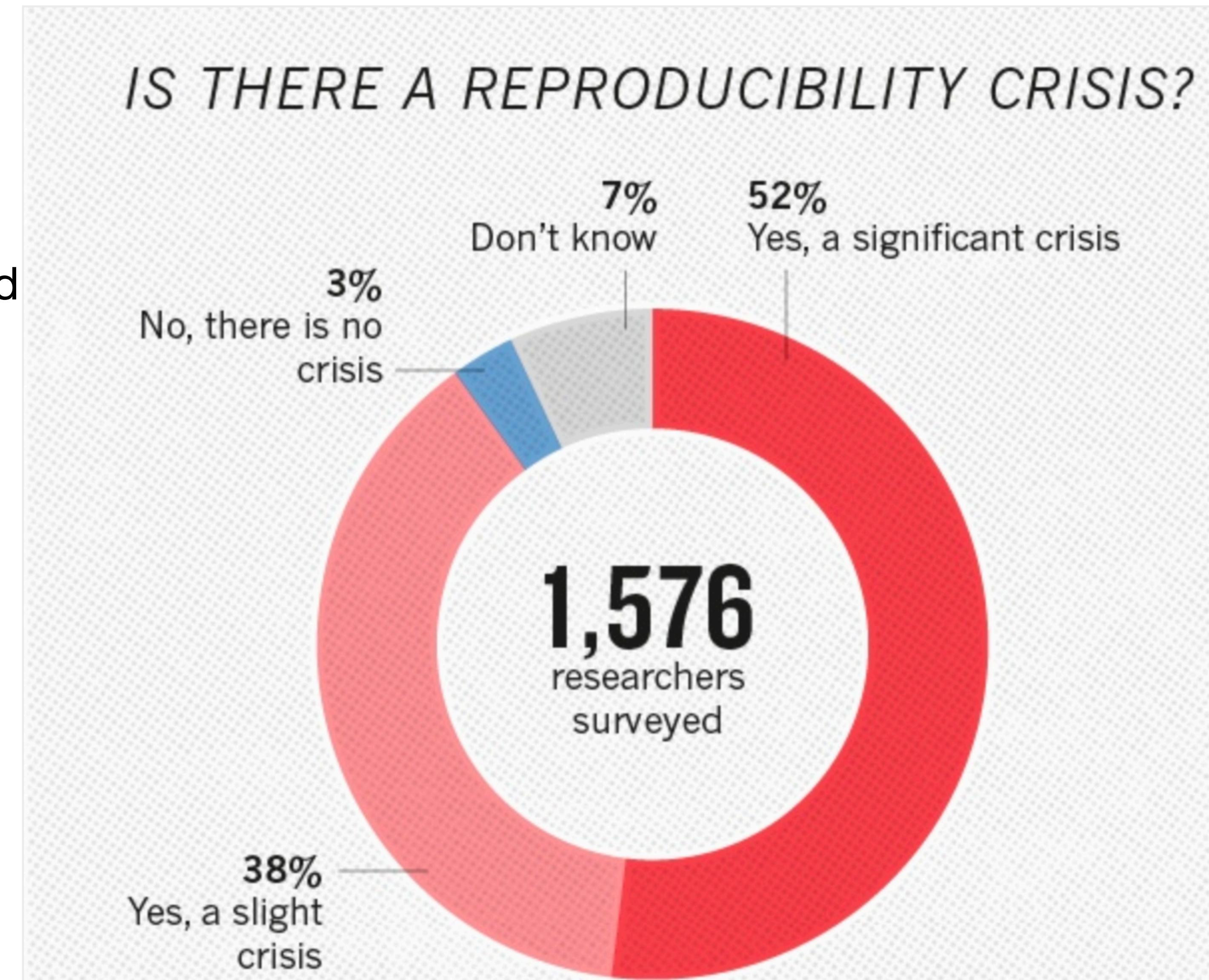
But how reproducible are most scientific findings?

Not very

Survey of >1500 researchers by Nature

- >70% of researchers have tried and failed to reproduce another lab's experiments
- >50% failed to reproduce their own experiments

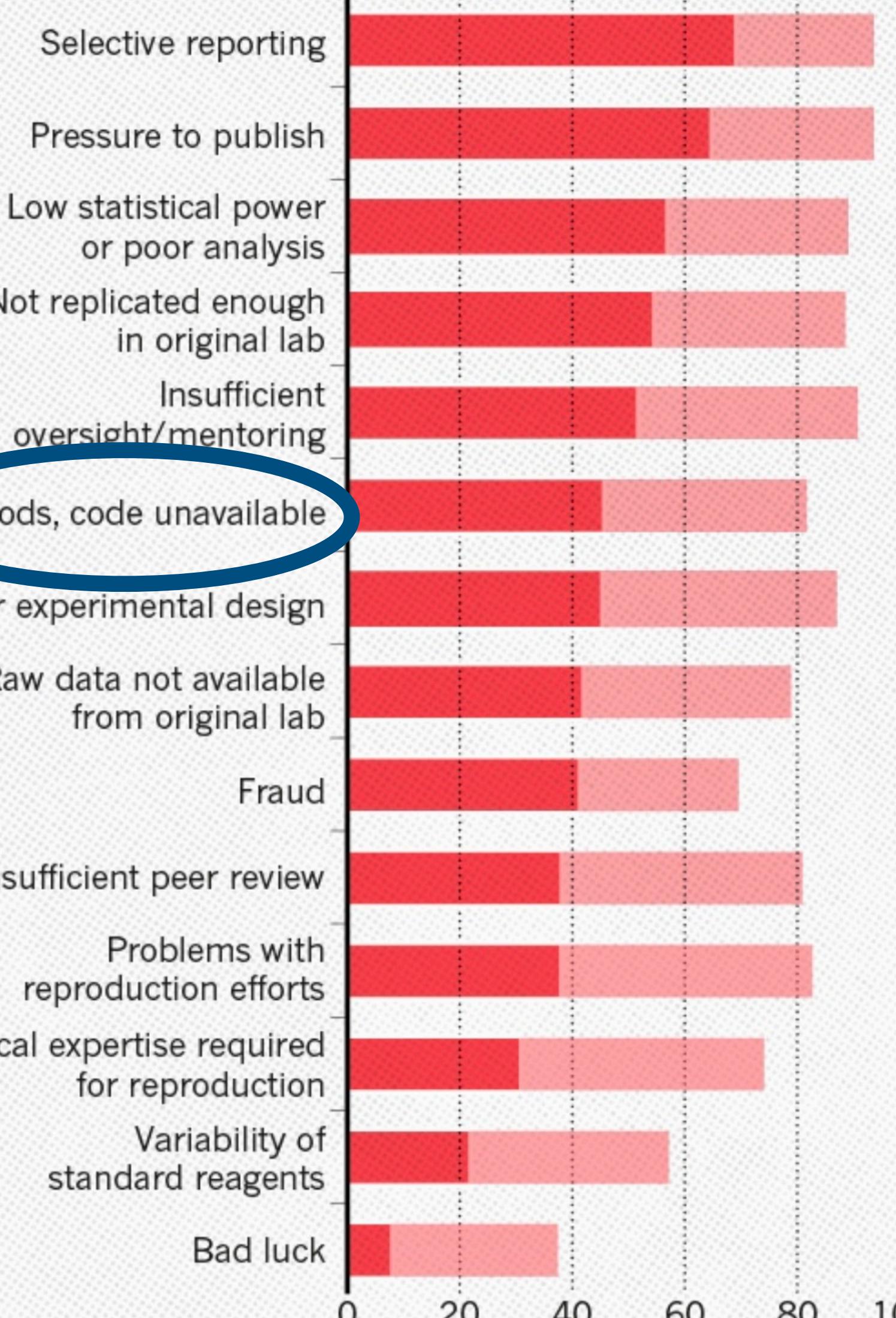
About half the respondents think there is a significant reproducibility crisis



WHAT FACTORS CONTRIBUTE TO IRREPRODUCIBLE RESEARCH?

Many top-rated factors relate to intense competition and time pressure.

- Always/often contribute
- Sometimes contribute



Why can't we reproduce our work

Major factors relate to some questionable activities

- selective reporting - p-hacking
- pressures to publish

Methods and code not available

- poorly described exact procedure can not be repeated

A similar problem has been found in other surveys and fields - this result has been replicated

The screenshot shows the Science journal website. At the top, there's a navigation bar with the AAAS logo, a "Become a Member" button, and links for "Contents", "News", "Careers", and "Journals". A red banner below the navigation bar says "Read our COVID-19 research and news.". The main content area features a research article titled "Estimating the reproducibility of psychological science" by the Open Science Collaboration. The article is marked as a "RESEARCH ARTICLE". Below the title, there are social sharing icons for Facebook, Twitter, LinkedIn, and Google+. The article summary includes a note about corresponding authors and affiliations, and a citation from Science magazine.

The screenshot shows the PLOS MEDICINE journal website. At the top, there's a navigation bar with links for "BROWSE", "PUBLISH", "ABOUT", and "PDF". The main content area features a research article titled "Why Most Published Research Findings Are False" by John P. A. Ioannidis. The article is marked as "OPEN ACCESS". Below the article, there's a section for "ESSAY".

Why Most Published Research Findings Are False

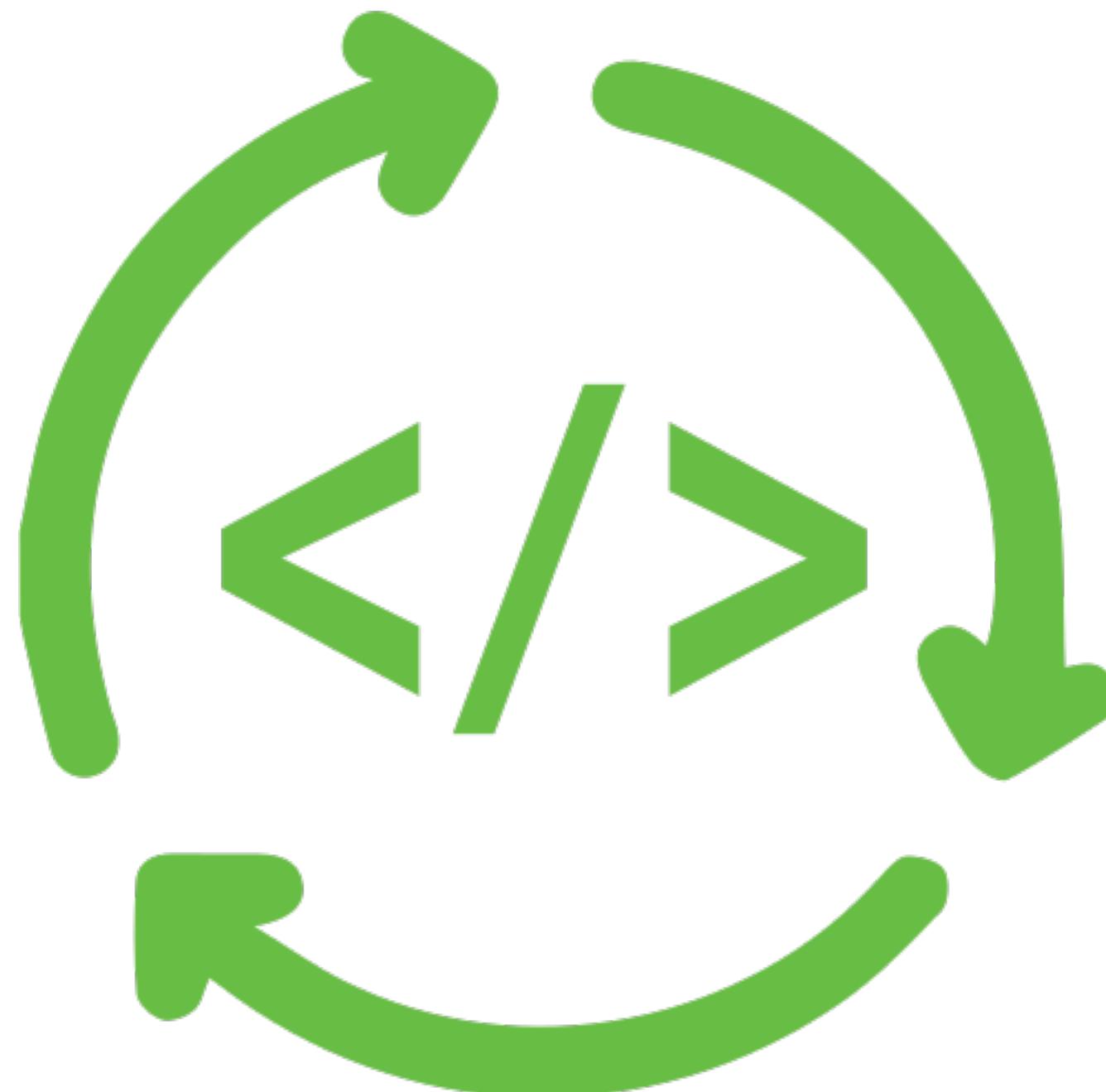
John P. A. Ioannidis

Published: August 30, 2005 • <https://doi.org/10.1371/journal.pmed.0020124>

Article	Authors	Metrics	Comments	Media Coverage
▼				

The screenshot shows TheScientist magazine website. At the top, there's a search bar and navigation links for "NEWS & OPINION", "MAGAZINE", and "SUBJ". The main content area features a news article titled "Half the Time, Psychology Results Not Reproducible: Study". The article discusses a study published on PsyArXiv that found half of psychological experiments fail to replicate. The author is Ashley P. Taylor, and the date is Nov 20, 2018. The article includes a quote from Brian Nosek. Below the article, there's a section for "EcoEvo Preprints" and a link to "Submit a Preprint".

To improve the reproducibility and reuse of the data we need to improve computational reproducibility



Dataset and analysis become more complex - accurately describing the analysis is key

Reviewers need access to data and code to be able to reproduce the results if necessary

Funding agencies and publishers require that research data be FAIR

Findable
Accessible
Interoperable
RReusable

Open Research Data



Research data should be freely accessible to everyone – for scientists as well as for the general public.

The SNSF agrees with this principle. Since October 2017, researchers have to include a data management plan (DMP) in their funding application for most of the funding schemes. At the same time, the SNSF expects that data generated by funded projects are publicly accessible in digital databases provided there are no legal, ethical, copyright or other issues.

Please consult the webpages of the different funding schemes to see whether a DMP is required when submitting an application.

SPRINGER NATURE

≡ Open research

About

Journals & books

Data

Institutional agreements

Funding & support

Publ



We believe that data should be open, accessible and reusable. Data sharing helps speed up the pace of discovery and its benefits to society.

As a proactive partner to the research community, Springer Nature is pioneering new approaches to data sharing and open data. We're committed to supporting researchers in sharing their data, helping to make data sharing the new normal.



ELSEVIER

About Elsevier

Products & Solutions

Services

Shop & Discover

Search

Home > Authors > Author resources > Research Data > Open Data

Open Data

What is open data?

Elsevier supports the [principle](#) ↗ that "Raw research data should be made freely available to all researchers" and authors should be free to publicly post their raw research data (see [Author Rights](#) for more details).

le

We have developed a simple way for authors to do this, by making their research data available on Mendeley Data and linking it to their article on ScienceDirect.

F1. (meta)data are assigned a globally unique and eternally persistent identifier.

F2. data are described with rich metadata.

F3. (meta)data are registered or indexed in a searchable resource.

F4. metadata specify the data identifier.

A1 (meta)data are retrievable by their identifier using a standardized communications protocol.

A1.1 the protocol is open, free, and universally implementable.

A1.2 the protocol allows for an authentication and authorization procedure, where necessary.

A2 metadata are accessible, even when the data are no longer available.

I1. (meta)data use a formal, accessible, shared, and broadly applicable language

I2. (meta)data use vocabularies that follow FAIR principles.

I3. (meta)data include qualified references to other (meta)data.

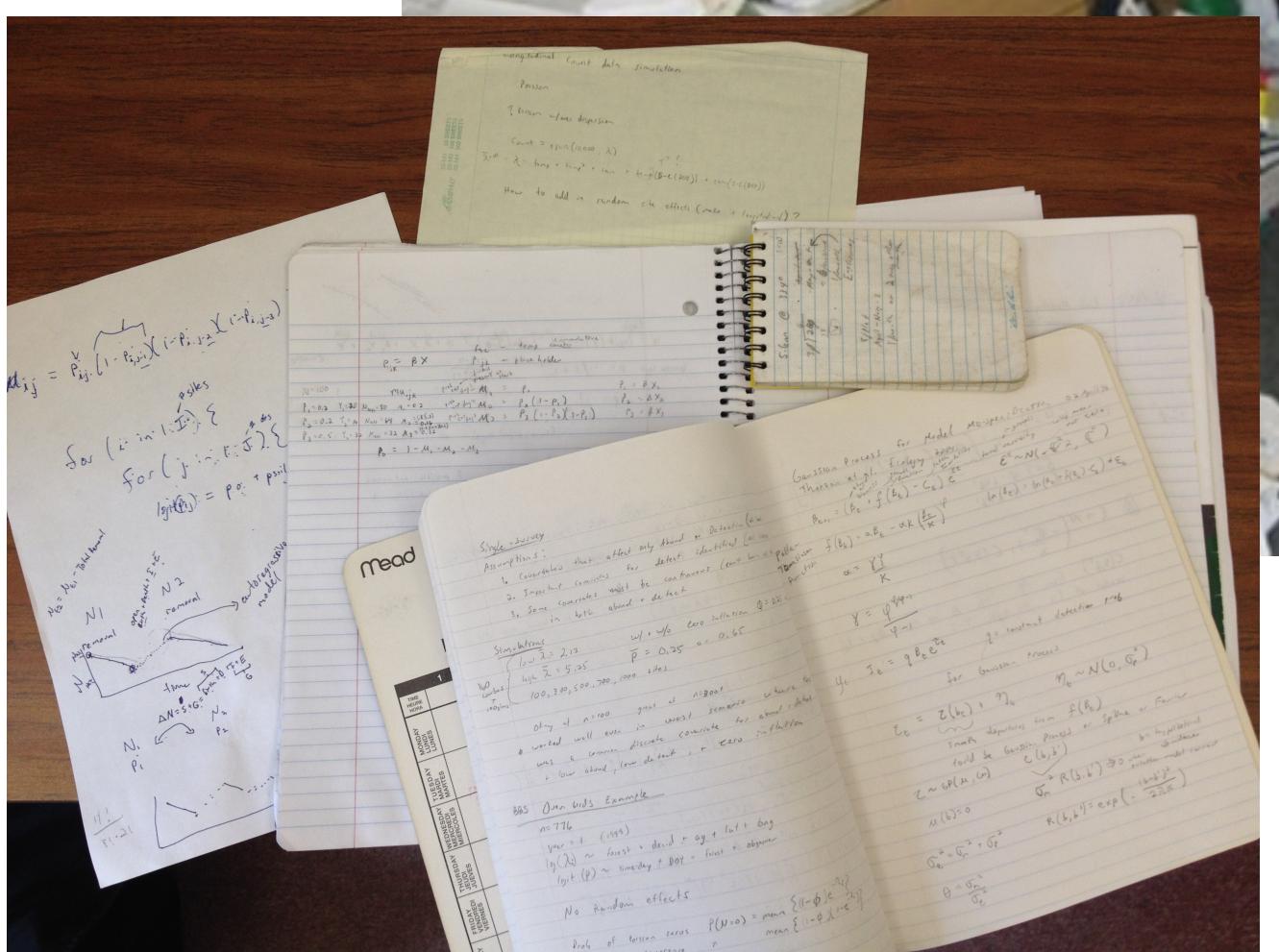
R1. meta(data) have a plurality of accurate and relevant attributes.

R1.1. (meta)data are released with a clear and accessible data usage license.

R1.2. (meta)data are associated with their provenance.

R1.3. (meta)data meet domain-relevant community standards.

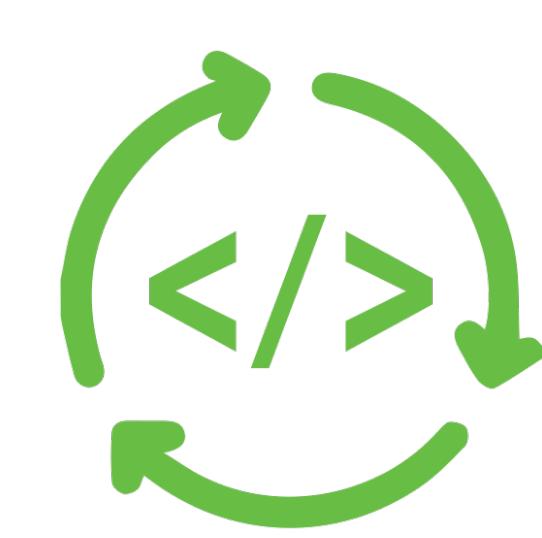
Being FAIR - starts with data good management



You need to know what you did
Record everything and write meticulous notes
Store your data safely - multiple copies

This is better





Data management and Computational reproducibility

- **Reproducible - by you or others**

Reproducible - by you

- The first person who will need to reproduce your results will probably be you
 - New data becomes available
 - Return to the project after some time
 - You give the project to student/collaborator
 - A reviewer suggests a change
 - You find an error and you don't know what went wrong

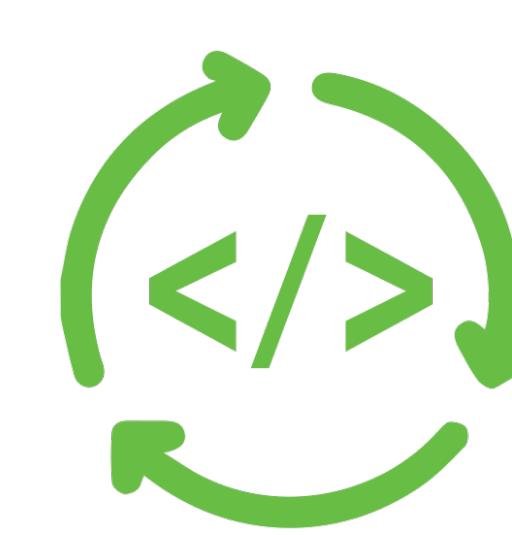




Data management and Computational reproducibility

- **Reproducible - by you or others**

“Someone unfamiliar with your project should be able to look at your computer files and understand in detail what you did and why” - Bill Noble



Data management and Computational reproducibility

- **Reproducible - by you or others**

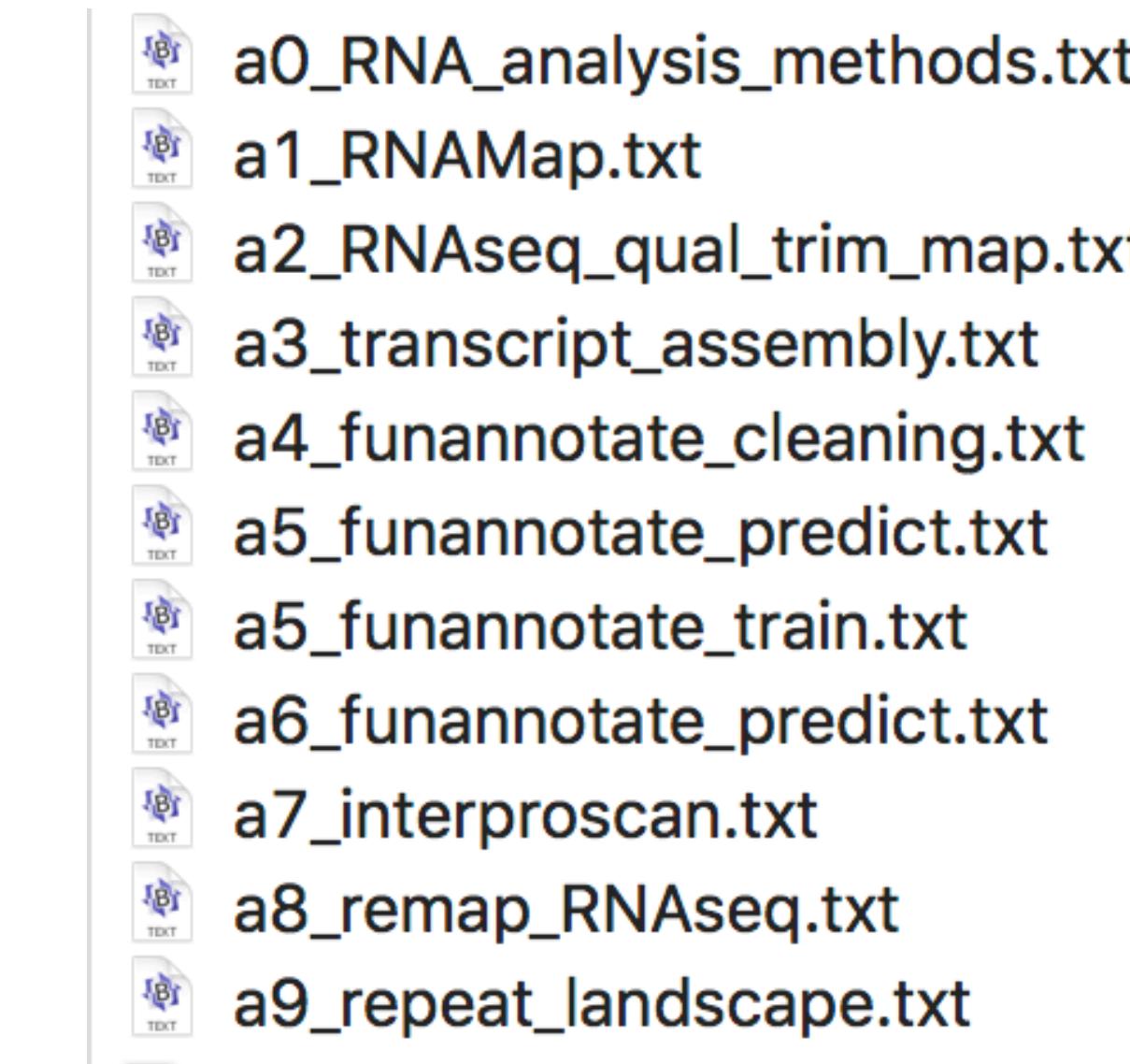
“Someone unfamiliar with your project should be able to look at your computer files and understand in detail what you did and why” - Bill Noble

- **Key components**
 - single well structured directory with meaningful subdirectories

Directory structure

```
project
|   README.md
|   data/
|       raw_data/
|           |   data_orig.csv
|       processed_data/
|           |   data_clean.csv
|       results/
|           |   model_results.csv
|   documents/
|       meeting_notes.md
|       data_dictionary.md
|   code/
|       exploration/
|           |   01_data_exploration.Rmd
|           |   02_model_results.Rmd
|       scripts/
|           |   01_do_clean_data.R
|           |   02_do_model_data.R
|       functions/
|           |   01_funs_clean_data.R
|           |   02_funs_model_data.R
```

Example of my own analysis scripts



 a0_RNA_analysis_methods.txt	
 a1_RNAMap.txt	
 a2_RNAseq_qual_trim_map.txt	
 a3_transcript_assembly.txt	
 a4_funannotate_cleaning.txt	
 a5_funannotate_predict.txt	
 a5_funannotate_train.txt	
 a6_funannotate_predict.txt	
 a7_interproscan.txt	
 a8_remap_RNAseq.txt	
 a9_repeat_landscape.txt	

Directory naming

Keep path names short (< 256 characters)

Recommendation for file names:

Unique, reflect content (if possible)

Use only ASCII (American Standard Code for Information Interchange) characters

- NO SPACES
- Be aware of case sensitivity

Bad examples

data%20management%20plan.docx
sup figure 2.png
lab meeting 19.10.2019.pptx

Good examples

Data_management_plan_SNF.docx
sup_figure_02_summary_stats.png
lab_meeting_2019-10-19.pptx



Data management and Computational reproducibility

- **Reproducible - by you or others**

“Someone unfamiliar with your project should be able to look at your computer files and understand in detail what you did and why” - Bill Noble

- **Key components**
 - single well structured directory with meaningful subdirectories
 - data processing and analysis should carefully documented in interactive notebooks based on open access software (e.g. Jupyter notebook, R markdown)

Data processing and analysis

- Interactive Notebooks - combine documentation, code, input and output generated by the code, e.g. graphs)
- Open access software - accessible
- Avoid WORD and EXCEL - versions, hidden characters, auto-filling, auto-formatting, scalability

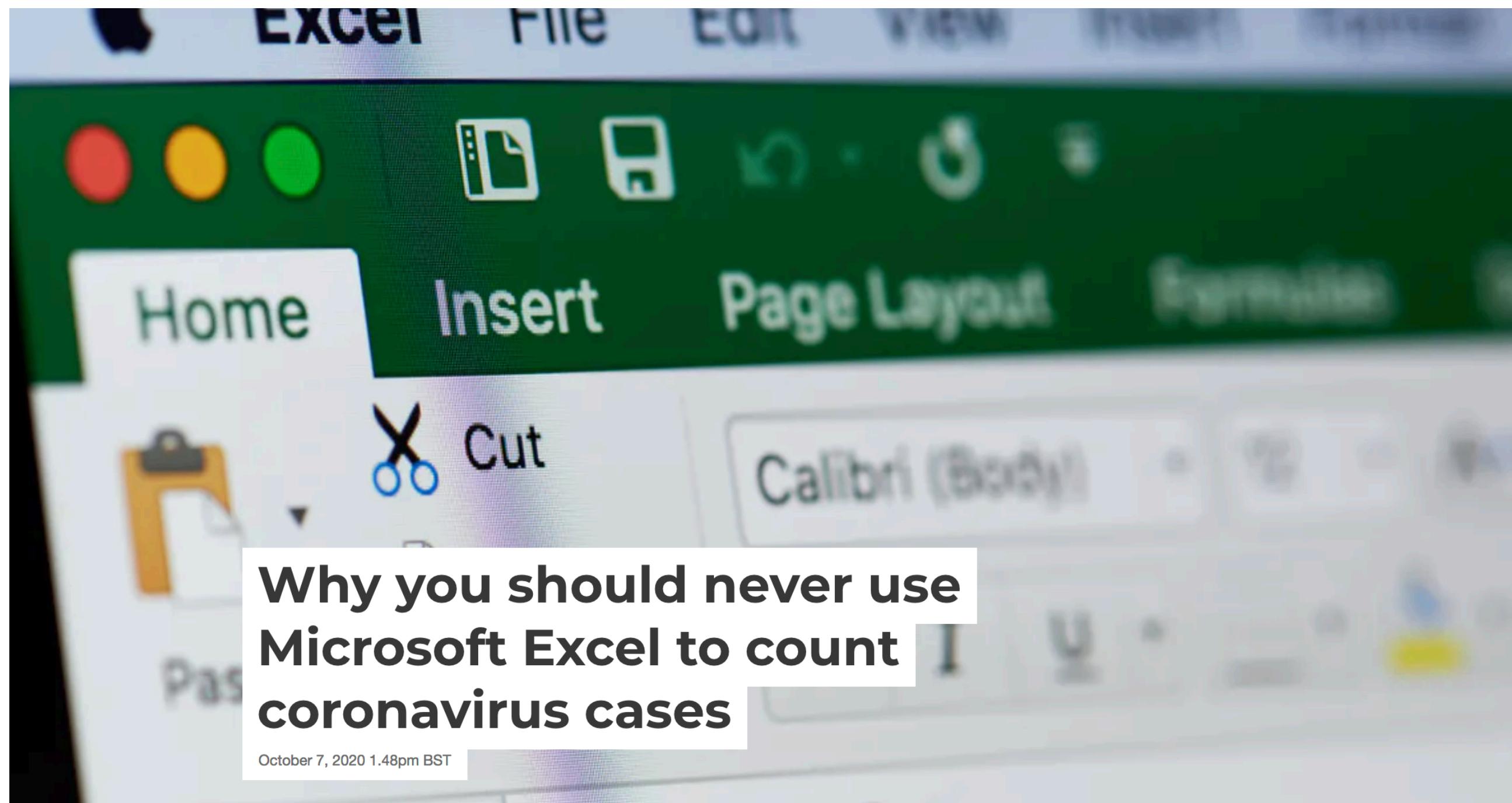


<https://rmarkdown.rstudio.com/>



<https://jupyter.org/>

It is OK, I have an excel spreadsheet with all my data



Tech

Excel: Why using Microsoft's tool caused Covid-19 results to be lost

By Leo Kelion
Technology desk editor

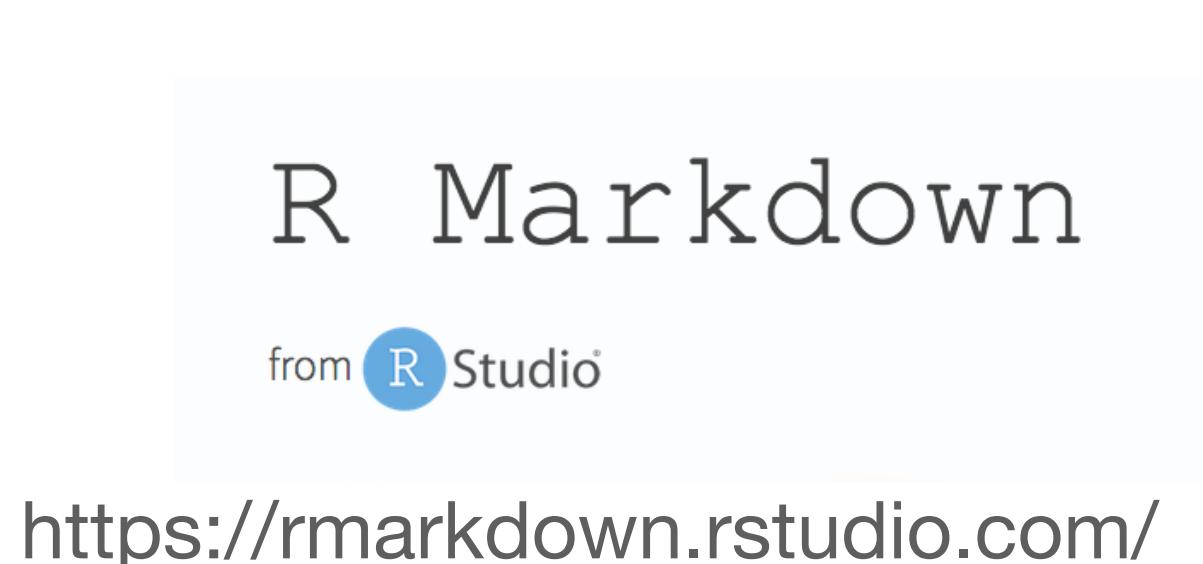
Covid: how Excel may have caused loss of 16,000 test results in England

Public Health England data error blamed on limitations of Microsoft spreadsheet

- 16000 coronavirus cases were missing from daily reports between 25.09-02.10.2020
- Underestimate the scale of infections
- Delayed Track and Trace - possibly increasing community transmission and risk of infection

Data processing and analysis

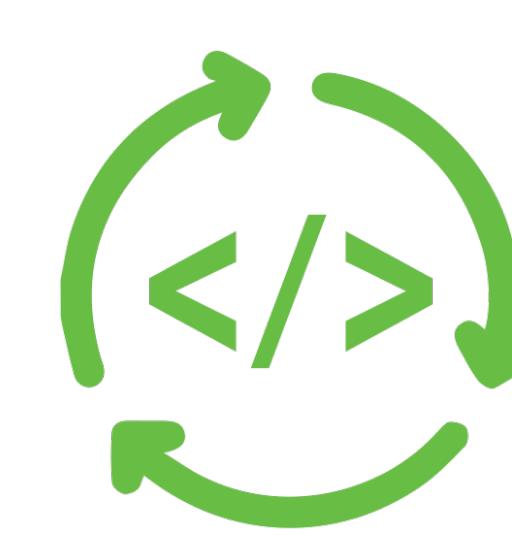
- Interactive Notebooks - combine documentation, code, input and output generated by the code, e.g. graphs)
- Open access software - accessible
- Avoid WORD and EXCEL - versions, hidden characters, auto-filling, auto-formatting, scalability
 - if you use excel - convert xls. files to files that are ‘open format’ not proprietary (e.g. csv, txt). Check the conversion of data, e.g. conversion to dates



Jessica Stapley



<https://jupyter.org/>



Data management and Computational reproducibility

- **Reproducible - by you or others**

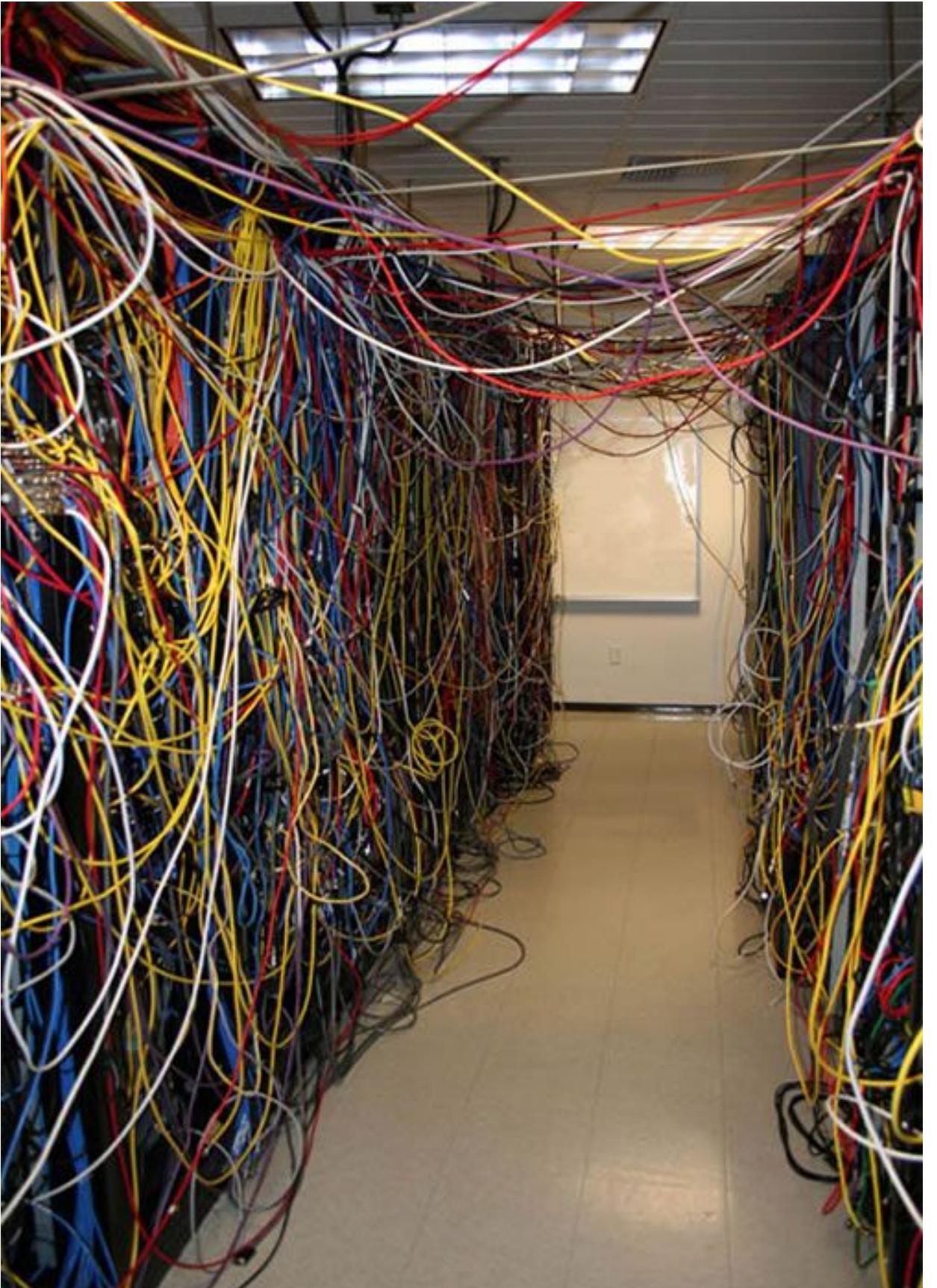
“Someone unfamiliar with your project should be able to look at your computer files and understand in detail what you did and why” - Bill Noble

- **Key components**

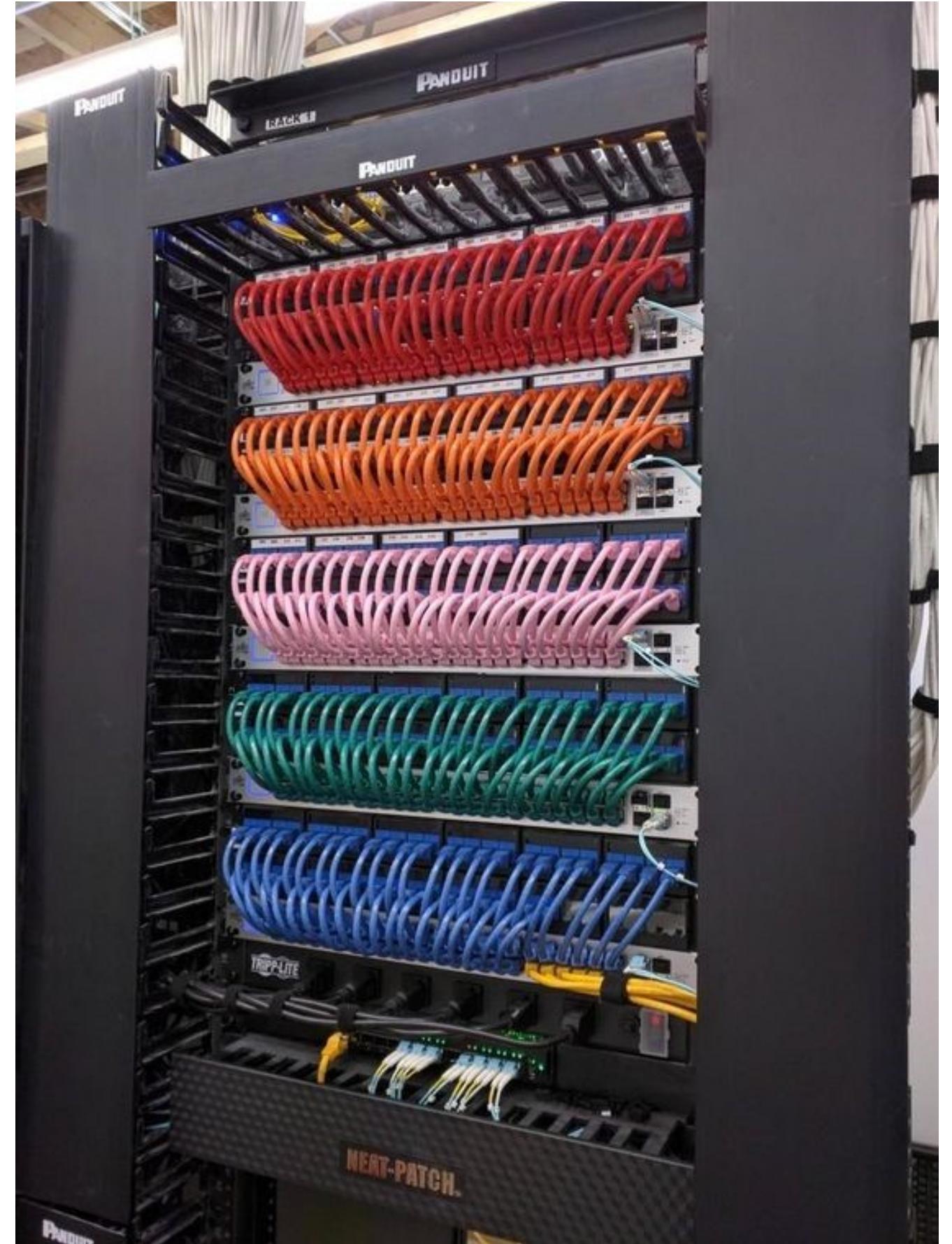
- single well structured directory with meaningful subdirectories
- data processing and analysis should carefully documented in interactive notebooks based on open access software (e.g. Jupyter notebook, R markdown)
- accessible and hosted on a repository (e.g. github)

Computational reproducibility

This is fine



This is better



Someone unfamiliar with your project should be able to look at your computer files and understand in detail what you did and why - Bill Noble



R projects and R markdown

Jessica Stapley



- RStudio is an integrated development environment (IDE) for R
- makes working with R easier - includes
 - drop down menus
 - syntax highlighting
 - code completion
 - smart indentation
 - workspace browser
 - data viewer
 - plot history
 - ... and much more

The screenshot shows the R Studio interface with several panels:

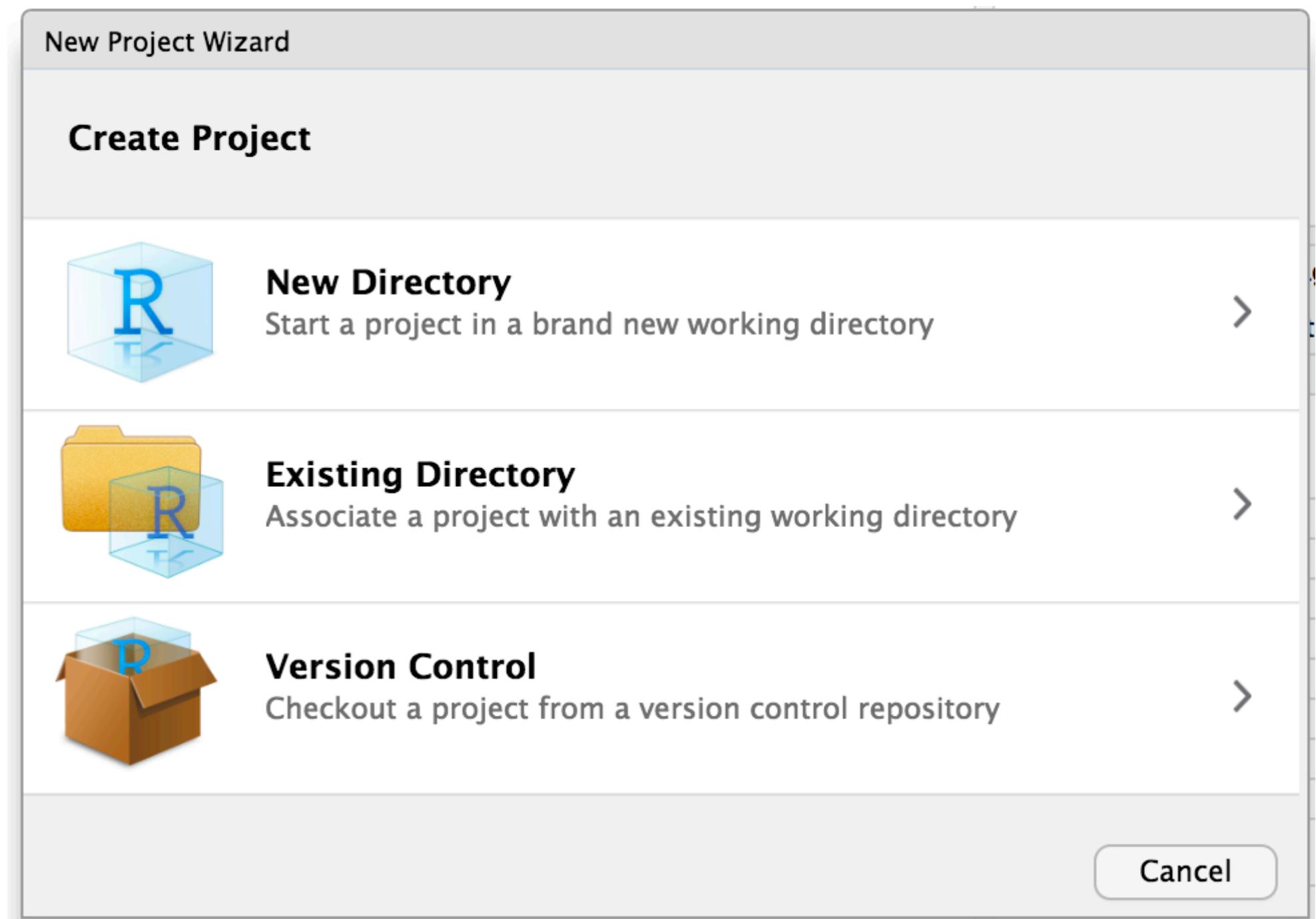
- Script Editor:** A large central panel titled "Untitled1" with the subtitle "R Script". It contains placeholder text: "This is where you write your script. You can run it, save it, and keep a record of your analysis."
- Console:** A panel at the bottom left showing R startup messages:

```
R version 3.4.3 (2017-11-30) -- "Kite-Eating Tree"  
Copyright (C) 2017 The R Foundation for Statistical Computing  
Platform: x86_64-w64-mingw32/x64 (64-bit)  
  
R is free software and comes with ABSOLUTELY NO WARRANTY.  
You are welcome to redistribute it under certain conditions.  
Type 'license()' or 'licence()' for distribution details.  
  
R is a collaborative project with many contributors.  
Type 'contributors()' for more information and  
'citation()' on how to cite R or R packages in publications.  
  
Type 'demo()' for some demos, 'help()' for on-line help, or  
'help.start()' for an HTML browser interface to help.  
Type 'q()' to quit R.
```
- Environment Browser:** A panel on the right titled "Global Environment" which is currently empty, indicated by the message "Environment is empty". It has tabs for "Environment", "History", "Connections", and "Git".
- File Browser:** A panel at the bottom right with tabs for "Files", "Plots", "Packages", "Help", and "Viewer".

Annotations provide descriptions for each panel:

- Script Editor:** "This is where you write your script. You can run it, save it, and keep a record of your analysis."
- Console:** "This is the console – code input and output can be seen here."
- Environment Browser:** "This shows information about your “environment” i.e. imported data and other objects in the workspace."
- File Browser:** "This panel has tabs to show the working directory, help files, package information and plots."

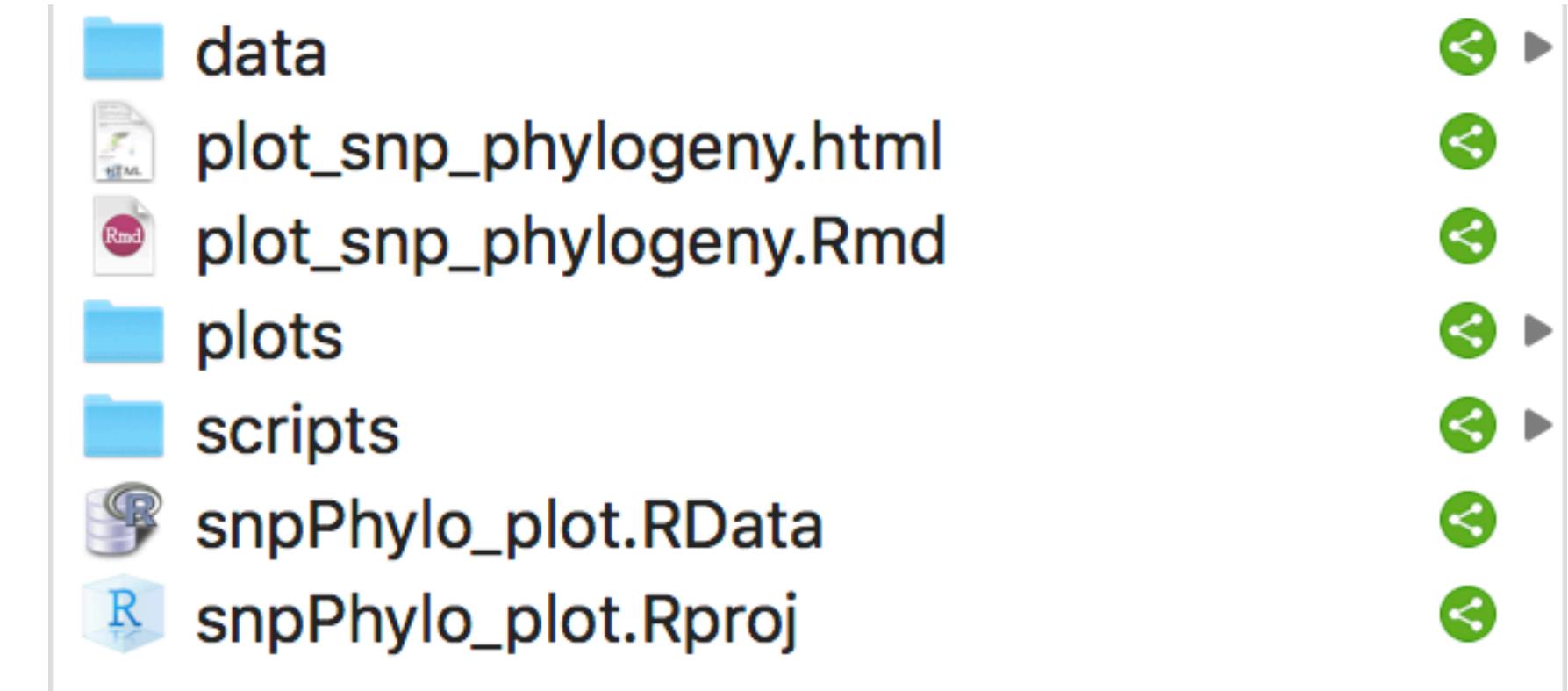
R studio - R projects and R Markdown



R projects

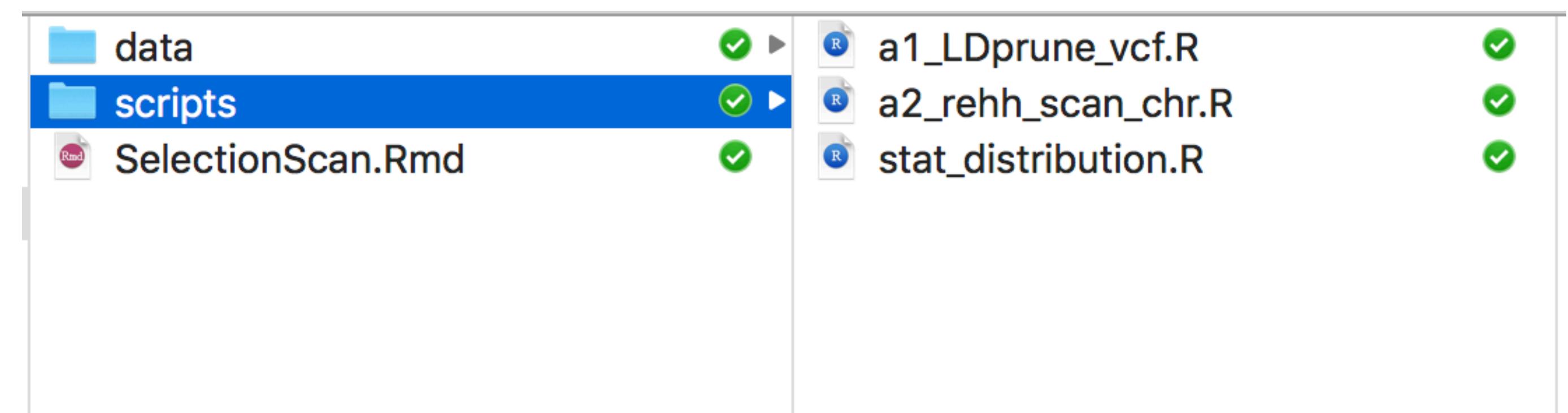
R projects

- creates a directory where all files and relevant metadata are kept
- keeps data reliable, portable, shareable and reproducible
- sets the environment (e.g. working directory) so the code will run on any machine

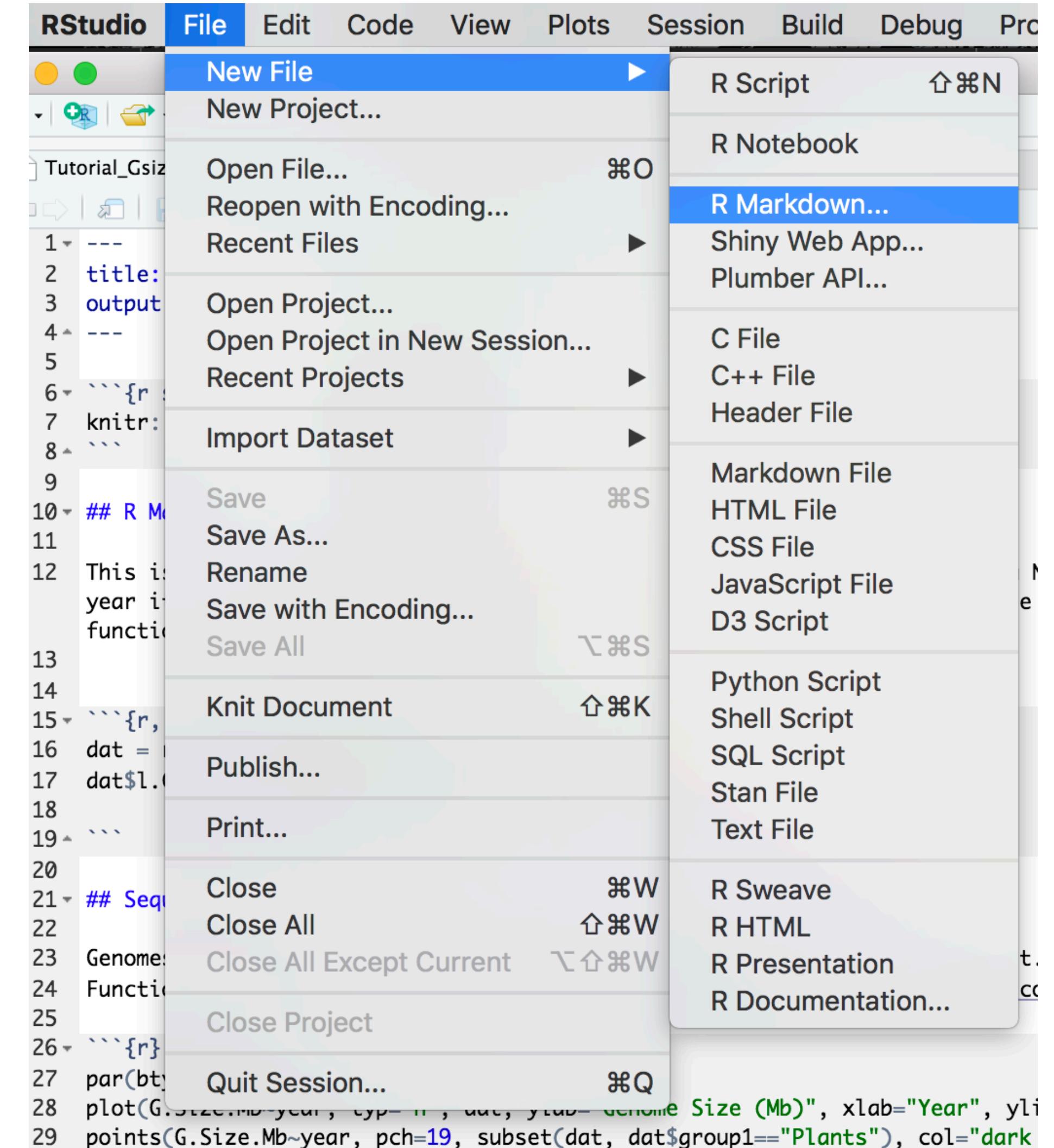
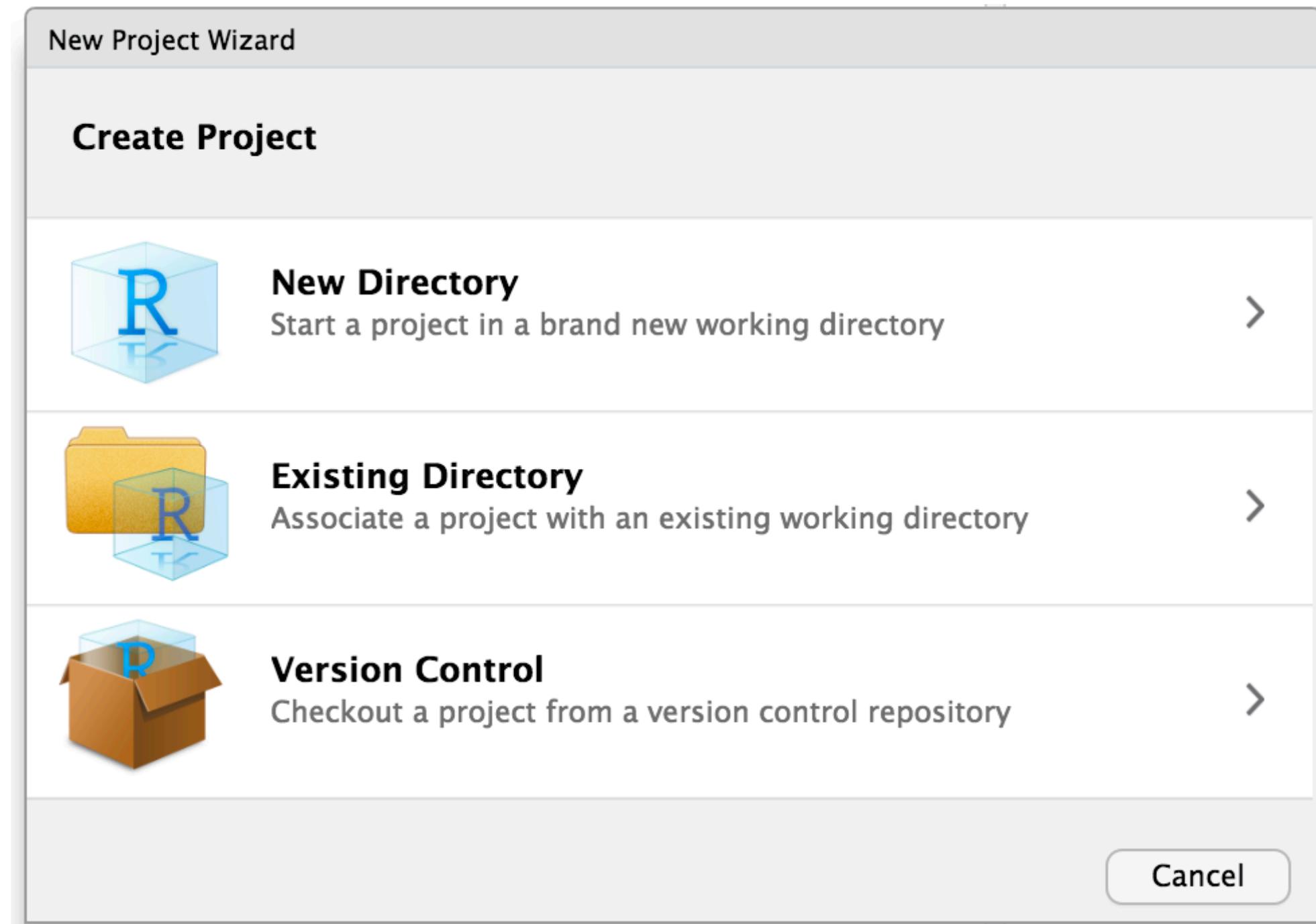


No single way to structure the project

- (raw)data: contains data for project (read only protected)
- R scripts - sequentially named so order can be easily seen
- plots: plots of the data



R studio - R projects and R Markdown



Markup and markdown

Markup - system for annotating text documents, e.g. html

```
<h1>Heading</h1>
<h2>Sub-heading</h2>
<a href="www.webpage.com">Link</a>
<ul>
  <li>List-item1</li>
  <li>List-item2</li>
  <li>List-item3</li>
</ul>
```

Markdown - lightweight markup language which uses plain-text syntax in order to be as unobtrusive as possible, so that a human can easily read it

R studio Markdown

```
---
```

```
title: "Genome size of sequenced genomes available on NCBI"
output: html_document
```

```
---
```

```
```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

```

```
## R Markdown
```

This is an R Markdown document to create plots of data on genome size from NCBI (as of 20.06.19). Here I plot genome size against the year it was released for all the genomes available on NCBI. I also plot the distribution of log genome size using a kernel density function.

```
```{r, echo=FALSE}
dat = read.table("data/NCBI_eukaryotes.txt", header = TRUE)
dat$l.Gsize = log(dat$G.Size.Mb)
```

```

```
## Sequenced genome size over time for different taxonomic groups
```

Genomes on NCBI have been getting bigger. The Axolotl genome is the largest.

Function to position and size the image from <https://scrogster.wordpress.com/2014/06/02/adding-phylopic-org-silhouettes-to-r-plots/>

```
```{r}
par(bty="l")
plot(G.Size.Mb~year, typ="n", dat, ylab="Genome Size (Mb)", xlab="Year", ylim=c(0,37000))
points(G.Size.Mb~year, pch=19, subset(dat, dat$group1=="Plants"), col="dark green")
points(G.Size.Mb~I(year+0.25), pch=19, subset(dat, dat$group1=="Animals"), col="blue")
points(G.Size.Mb~I(year+0.5), pch=19, subset(dat, dat$group1=="Fungi"), col="orange")
points(G.Size.Mb~I(year+0.75), pch=19, subset(dat, dat$group1=="Protists"), col="red")
```

```

Genome size of sequenced genomes available on NCBI

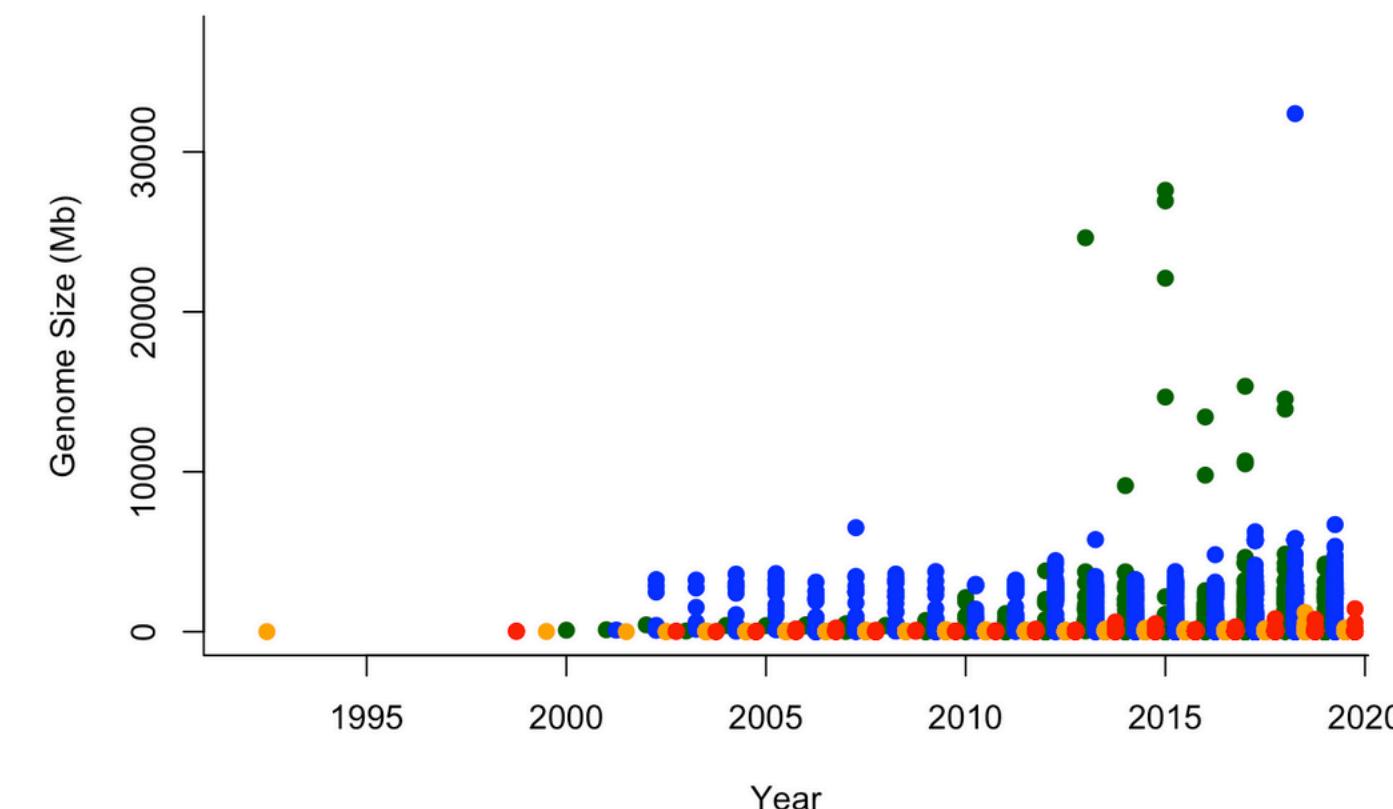
R Markdown

This is an R Markdown document to create plots of data on genome size from NCBI (as of 20.06.19). Here I plot genome size against the year it was released for all the genomes available on NCBI. I also plot the distribution of log genome size using a kernel density function.

Sequenced genome size over time for different taxonomic groups

Genomes on NCBI have been getting bigger. The Axolotl genome is the largest. Function to position and size the image from <https://scrogster.wordpress.com/2014/06/02/adding-phylopic-org-silhouettes-to-r-plots/>

```
par(bty="l")
plot(G.Size.Mb~year, typ="n", dat, ylab="Genome Size (Mb)", xlab="Year", ylim=c(0,37000))
points(G.Size.Mb~year, pch=19, subset(dat, dat$group1=="Plants"), col="dark green")
points(G.Size.Mb~I(year+0.25), pch=19, subset(dat, dat$group1=="Animals"), col="blue")
points(G.Size.Mb~I(year+0.5), pch=19, subset(dat, dat$group1=="Fungi"), col="orange")
points(G.Size.Mb~I(year+0.75), pch=19, subset(dat, dat$group1=="Protists"), col="red")
```



Integration of GitHub and R

- Advanced users
- Happy git with R (<https://happygitwithr.com/>)
- Git is a version control system developed to help groups of people work together collaboratively on software projects
- Github is a web-based hosting service

Summary

- We need to ensure our research is reproducible
- FAIR - guide to enhance the reusability of their data
- Recommended (required) by publishers and funding bodies
- Assess the FAIRness of your data ([https://ardc.edu.au/resources/working-with-data/
fair-data/fair-self-assessment-tool/](https://ardc.edu.au/resources/working-with-data/fair-data/fair-self-assessment-tool/))
- It all starts with good research data management and clear description of what you did and what versions of software you used.
- R studio, R projects, R markdown and integration with Github can help you make your data more open, shareable and reproducible