Reproducible Research

Marco Chiapello

27/03/2017

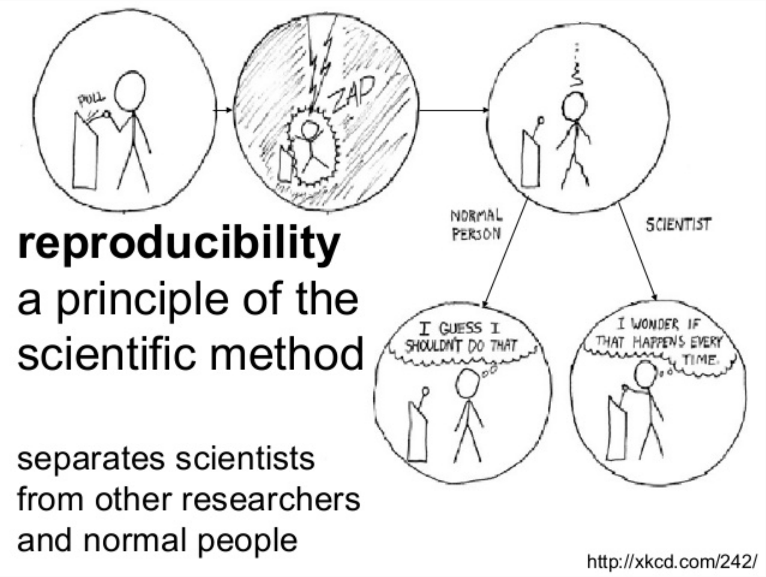
Center for Proteomics
University of Cambridge
mc983@cam.ac.uk

Overview

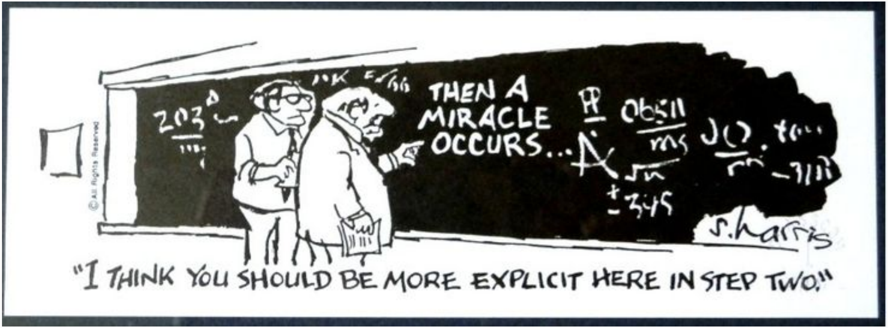
Introduction

Replication is the ultimate standard by which scientific claims are judged [2]
The fact that an analysis is **reproducible does not guarantee the quality, correctness, or validity** of the published results.

What reproducible research is



What reproducible research is



- This is exactly how it seems when you try to figure out how authors got from a large and complex data set to a dense paper with lots of busy figures.
  Without access to the **data and the analysis code**, a miracle occurred.
- And there should be NO MIRACLES IN SCIENCE. [1]

$$\mathit{DATA} + \mathit{ANALYSIS} \rightarrow \mathit{RESULTS}$$

Common practice of writing statistical reports:

- We import a dataset into Excel
- Run a procedure to get the results
- Copy and paste selected pieces into a typesetting program (e.g. Word)
- Add a few descriptions
- Finish a report

There are obvious dangers and disadvantages in this process:

1. It is **error-prone** due to too much manual work;
2. It requires lots of human effort to do **tedious jobs**;
3. The workflow is barely recordable, therefore it is **difficult to reproduce**;
4. A **tiny change** of the data source in the future will require the author(s) to go through the same procedure again;
5. The analysis and writing are separate, so close attention has to be paid to the **synchronization of the two parts**.

What is Reproducible Research?

The ability to reproduce someone else results

What do you need?

- Analytic data
- Analytic code
- **Documentation for data and code**

What reproducible research is

Reproducible vs Replicable

		DATA	
		Same	Different
CODE	Same	Reproducible	Replicable
	Different	Robust	Generalisable

Ref: `https://github.com/KirstieJane/ReproducibleResearch`

What reproducible research is

Reproducibility/reproduce

A study is reproducible if there is a specific set of computational functions/analyses (usually specified in terms of code) that **exactly reproduce all of the numbers in a published paper from raw data**.

Replication/replicate

A study is only replicable if you perform the exact same experiment (at least) twice, collect data in the same way both times, perform the same data analysis, and **arrive at the same conclusions**.

Reproducibility is, to some extent, a technical challenge, while **Replication** gives to the results scientific validity.

Ref: `https://github.com/lgatto/TeachingMaterial/tree/master/open-rr-bioinfo-best-practice`

Reproducible research Reasons

How does working reproducibly help to achieve more as a scientist [1]

Reasons to work reproducibly

REPRODUCIBILITY [1]

Idealist:

1. It is the foundation of science!
2. The world would be a better place if everyone worked transparently and reproducibly!

Reasons to work reproducibly

REPRODUCIBILITY [1]

Idealist:

1. It is the foundation of science!
2. The world would be a better place if everyone worked transparently and reproducibly!

Realist:

1. It helps to avoid disaster
    - You need to record in detail how you got there
    - Work reproducibly early on will save you time later

REPRODUCIBILITY [1]

Idealist:

1. It is the foundation of science!
2. The world would be a better place if everyone worked transparently and reproducibly!

Realist:

1. It helps to avoid disaster
    - You need to record in detail how you got there
    - Work reproducibly early on will save you time later
2. It makes it easier to write papers
    - To have very transparent data and code, it costs just few minutes to spot a mistake (if any)

REPRODUCIBILITY [1]

Idealist:

1. It is the foundation of science!
2. The world would be a better place if everyone worked transparently and reproducibly!

Realist:

1. It helps to avoid disaster
    - You need to record in detail how you got there
    - Work reproducibly early on will save you time later
2. It makes it easier to write papers
    - To have very transparent data and code, it costs just few minutes to spot a mistake (if any)
3. It helps reviewers see it your way
    - Made the data and well-documented code easily accessible to the reviewers

REPRODUCIBILITY [1]

Idealist:

1. It is the foundation of science!
2. The world would be a better place if everyone worked transparently and reproducibly!

Realist:

1. It helps to avoid disaster
    - You need to record in detail how you got there
    - Work reproducibly early on will save you time later
2. It makes it easier to write papers
    - To have very transparent data and code, it costs just few minutes to spot a mistake (if any)
3. It helps reviewers see it your way
    - Made the data and well-documented code easily accessible to the reviewers
4. It enables continuity of your work
    - How can you ensure the continuity of work in your lab if progress is not documented reproducibly?
    - No proof of reproducibility, no result!

REPRODUCIBILITY [1]

Idealist:

1. It is the foundation of science!
2. The world would be a better place if everyone worked transparently and reproducibly!

Realist:

1. It helps to avoid disaster
   - You need to record in detail how you got there
   - Work reproducibly early on will save you time later
2. It makes it easier to write papers
   - To have very transparent data and code, it costs just few minutes to spot a mistake (if any)
3. It helps reviewers see it your way
   - Made the data and well-documented code easily accessible to the reviewers
4. It enables continuity of your work
   - How can you ensure the continuity of work in your lab if progress is not documented reproducibly?
   - No proof of reproducibility, no result!
5. It helps to build your reputation
   - To build a reputation for being an honest and careful researcher

Reproducible research Rules

– based on Sandve et al., 2013 [3]

Rule 1

For Every Result, Keep Track of How It Was Produced

Rule 1

For Every Result, Keep Track of How It Was Produced

- The **full sequence** of pre- and post-processing steps are often critical in order to reach the achieved result
- **Every detail** that may influence the execution of the step **should be recorded**
- Include the name and version of the program, as well as the exact parameters and inputs

As a minimum, you should at least record sufficient details on programs, parameters, and manual procedures to allow yourself, in a year or so, to approximately reproduce the results

Rule 2

Avoid Manual Data Manipulation Steps

Rule 2

Avoid Manual Data Manipulation Steps

- Manual procedures are not only inefficient and error-prone, they are also difficult to reproduce
- Manual modification of files can usually be replaced by the use of standard UNIX commands or scripts
- Manual tweaking of data files to attain format compatibility should be replaced by format converters that can be reenacted and included into executable workflows
- Manual operations like the use of **copy and paste** between documents should also be avoided

If manual operations cannot be avoided, you should as a minimum note down which data files were modified or moved, and for what purpose

Rule 3

Archive the Exact Versions of All External Programs Used

Rule 3

Archive the Exact Versions of All External Programs Used

- In order to exactly reproduce a given result, it may be necessary to use programs in the **exact versions used originally**
- It is not always trivial to get hold of a program in anything but the current version

As a minimum, you should note the exact names and versions of the main programs you use

Rule 4

Version Control All Custom Scripts

Rule 4

Version Control All Custom Scripts

- **Only that exact state of the script may be able to produce that exact output**, even given the same input data and parameters
- The standard solution to track evolution of code is to use a version control system
  - A version control system is a repository of files with monitored access. *Every change made to the source is tracked, along with who made the change, why they made it*

As a minimum, you should archive copies of your scripts from time to time

Rule 5

Record All Intermediate Results, When Possible in Standardized Formats

Rule 5

Record All Intermediate Results, When Possible in Standardized Formats

- In principle, as long as the **full process** used to produce a given result is tracked, all **intermediate data can also be regenerated**
- In practice, having easily **accessible intermediate results** may be of great value
- When the full process is not readily executable, it allows parts of the process to be rerun
- It **allows critical examination** of the full process behind a result

As a minimum, archive any intermediate result files that are produced when running an analysis

Rule 6

For Analyses That Include Randomness, Note Underlying Random Seeds

Rule 6

For Analyses That Include Randomness, Note Underlying Random Seeds

- Many analyses and predictions include some element of randomness, meaning the same program will typically give **slightly different results** every time it is executed
- Given the **same initial seed**, all random numbers used in an analysis will be equal, thus giving identical results every time it is run

As a minimum, you should note which analysis steps involve randomness, so that a certain level of discrepancy can be anticipated when reproducing the results

Rule 7

Always Store Raw Data

Rule 7

Always Store Raw Data

- Always store in a safe place the raw data
- Never touch or mofidy the raw data

Rule 8

Generate Hierarchical Analysis Output, Allowing Layers of Increasing Detail to Be Inspected

Rule 8

Generate Hierarchical Analysis Output, Allowing Layers of Increasing Detail to Be Inspected

- The final results that make it to an article, be it plots or tables, often represent **highly summarized data**
- In order to validate and fully understand the main result, it is often useful to inspect the detailed **values underlying the summaries**
- When working with summarized results, you should as a minimum at least once generate, inspect, and validate the detailed values underlying the summaries

Rule 9

Connect Textual Statements to Underlying Results

Rule 9

Connect Textual Statements to Underlying Results

- The results of analyses and their corresponding textual interpretations are clearly interconnected but often **lie in different places**
- Results usually live on a personal computer, while interpretations live in text documents
- To allow efficient retrieval of details behind textual statements, we suggest that **statements are connected to underlying results** already from the time the statements are initially formulated
- **Integrate reproducible analyses directly into textual documents**

Rule 10

Provide Public Access to Scripts, Runs, and Results

Rule 10

Provide Public Access to Scripts, Runs, and Results

- All input data, scripts, versions, parameters, and inter-mediate **results should be made publicly and easily accessible**
- Making reproducibility of your work by peers a realistic possibility sends a **strong signal of quality, trustworthiness, and transparency**

Reproducible research Tools

Let us change our traditional attitude to the construction of programs: Instead of imagining that our main task is to instruct a computer what to do, let us concentrate rather on **explaining to humans what we want the computer to do***.*

– Donald E. Knuth Literate Programming, 1984

Folder organization

```
project
|- doc/              # documentation for the study
|  +- paper/         # manuscript(s), whether generated or not
|
|- data              # raw and primary data, are not changed once created
|  |- raw/           # raw data, will not be altered
|  +- clean/         # cleaned data, will not be altered once created
|
|- code/             # any programmatic code
|- results           # all output from workflows and analyses
|  |- figures/       # graphs, likely designated for manuscript figures
|  +- pictures/      # diagrams, images, and other non-graph graphics
|
|- scratch/          # temporary files that can be safely deleted or lost
|- README            # the top level description of content
|- study.Rmd         # executable Rmarkdown for this study, if applicable
|- Makefile          # executable Makefile for this study, if applicable
|- study.Rproj       # RStudio project for this study, if applicable
|- datapackage.json # metadata for the (input and output) data files
```

Tools

A freely available language and environment for programming, computing and graphics

In this course we will use:

- `UNIX`
- `R`

Tools

Literate programming is a methodology that combines a programming language with a documentation language

- Write program code to do computing
- Write narratives to explain what is being done by the program code

Tools

Ref: `http://rmarkdown.rstudio.com/index.html`

RMarkdown

```
---
title: "Untitled"
author: "Marco Chiapello"
date: "10 June 2016"
output: html_document
---

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word
documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the
output of any embedded R code chunks within the document. You can embed an R code chunk like this:

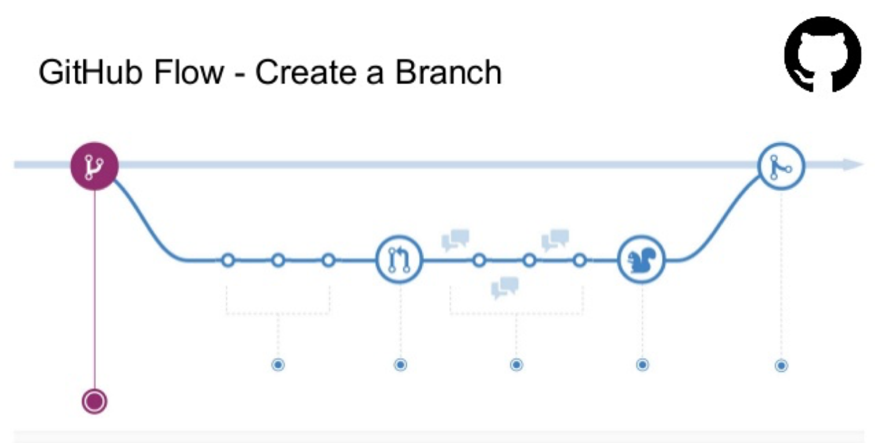
```{r}
summary(cars)
```

You can also embed plots, for example:

```{r, echo=FALSE}
plot(cars)
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that
generated the plot.
```

VERSION CONTROL



Tools

Learning to use these tools will require **commitment** and a **massive investment of your time and energy**.

A priori it is not clear why the benefits of working reproducibly outweigh its costs.

Does reproducibility sound like extra work?

It can be, particularly when one is first trying to do it, that is, to break one's own previous nonreproducible habits

Conclusion

My advice is:

Learn the tools of reproducibility as quickly as possible and use them in every project.

References

[1] Markowetz, F. (2016). Five selfish reasons to work reproducibly. *Genome biology*, pages 1–4.

[2] Peng, R. D. (2011). Reproducible research in computational science. *Science (New York, NY)*, 334(6060):1226–1227.

[3] Sandve, G. K., Nekrutenko, A., Taylor, J., and Hovig, E. (2013). Ten Simple Rules for Reproducible Computational Research. *PLoS Computational Biology*, 9(10):e1003285–4.