Reproducible Research

Marco Chiapello

June 9, 2016

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

Overview

Introduction

Reproducible research Reasons

Reproducible research Tools

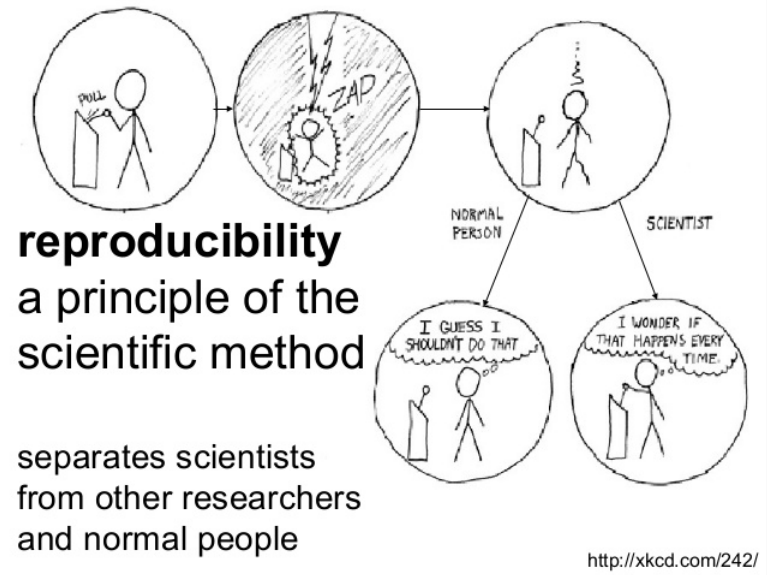
Reproducible research Rules

Conclusion

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



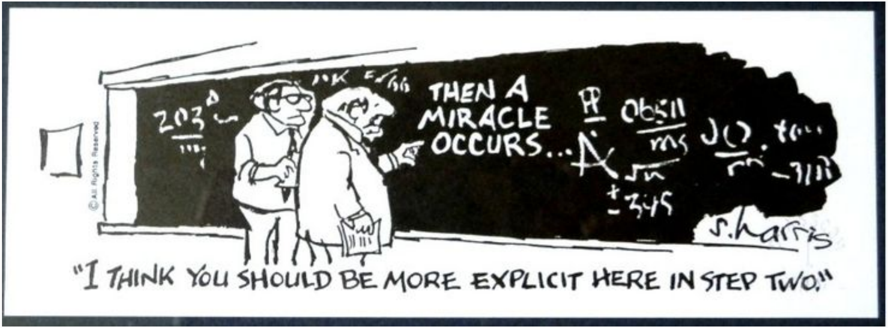
reproducibility
a principle of the scientific method

separates scientists from other researchers and normal people

http://xkcd.com/242/

PULL
ZAP
NORMAL PERSON
SCIENTIST
I GUESS I SHOULDN'T DO THAT
I WONDER IF THAT HAPPENS EVERY TIME.

What reproducible research is



"I THINK YOU SHOULD BE MORE EXPLICIT HERE IN STEP TWO."

- 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]

What reproducible research is

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

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.

Replicability requires new samples and new data[1], which introduces new variability, and additional risks of errors. **Reproducibility** is, to some extent, a technical challenge, while replication gives the results scientific validity.

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

[1] in particular biological replicates

Reproducible research Reasons

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

Reasons to work reproducibly

REPRODUCIBILITY

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 pro- gress 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 Tools

Tools

Literate programming is a methodology that combines a programming language with a documentation language, thereby making programs more robust, more portable, more easily maintained, and arguably more fun to write than programs that are written only in a high-level language.

Tools

At the lowest level, working reproducibly just means avoiding beginners? mistakes. Keep your project organized, name your files and directories in some informative way, store your data and code at a single backed-up location. Don?t spread your data over different servers, laptops and hard drives. To achieve the next levels of reproducibility, you need to learn some tools of computational reproducibility [8]. In general, reproducibility is improved when there is less clicking and pasting and more scripting and coding. For example, do your analysis in R (https://www.r-project.org/) or Python (https://www.python.org/) and document your analysis using knitR (http://yihui.name/knitr/) or IPython notebooks (http://ipython.org/). These tools help you to merge descriptive text with analysis code into dynamic documents that can be automatically updated every time the data or code change. As a next step, learn

Tools

Learning the tools of the trade 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

Reproducible research Rules

Rule 1

Rule 2

Rule 3

Rule 4

Rule 5

Rule 6

Rule 7

Rule 8

Rule 9

Rule 10

Conclusion

Conclusions

My advice is: learn the tools of reproducibility (Box 1) 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.

Acknowledgements

- Mike Deery Manager of the Cambridge Center for Proteomics
- Laurent Gatto Computational Proteomics Unit

- Andrea Genre Universita' di Torino