# Reproducible Research in Bioinformatics and Data Mining

Adrienn Szabó

DMS Group, MTA SZTAKI

October 2, 2014

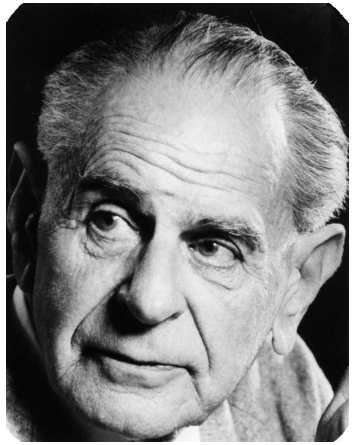
# What is (not) Reproducible Research?


Courtesy of S. Harris, ScienceCartoonsPlus.com.

# What is (not) Reproducible Research?

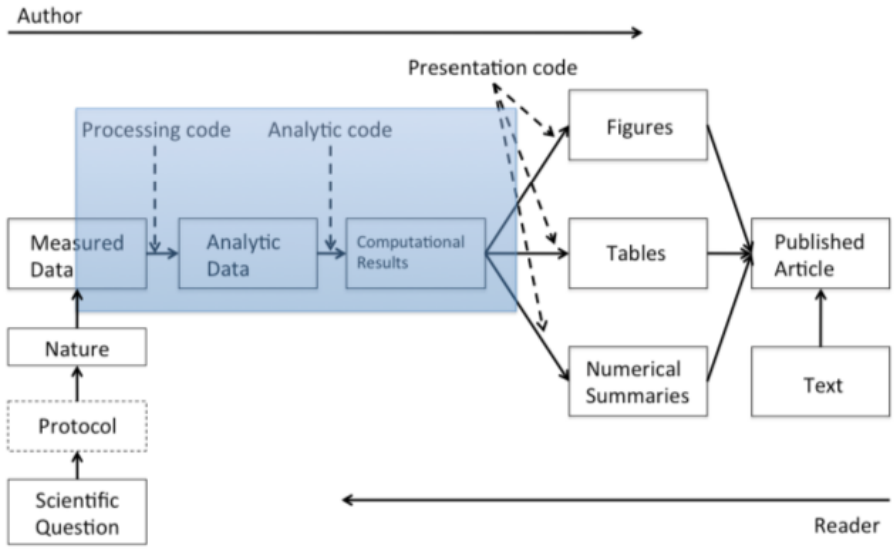
If you observe (or measure, simulate) something but it's not repeatable or reproducible, then it's NO science.

> "***. . . non-reproducible single occurrences are of no significance to science.***"
>
> — Karl Popper

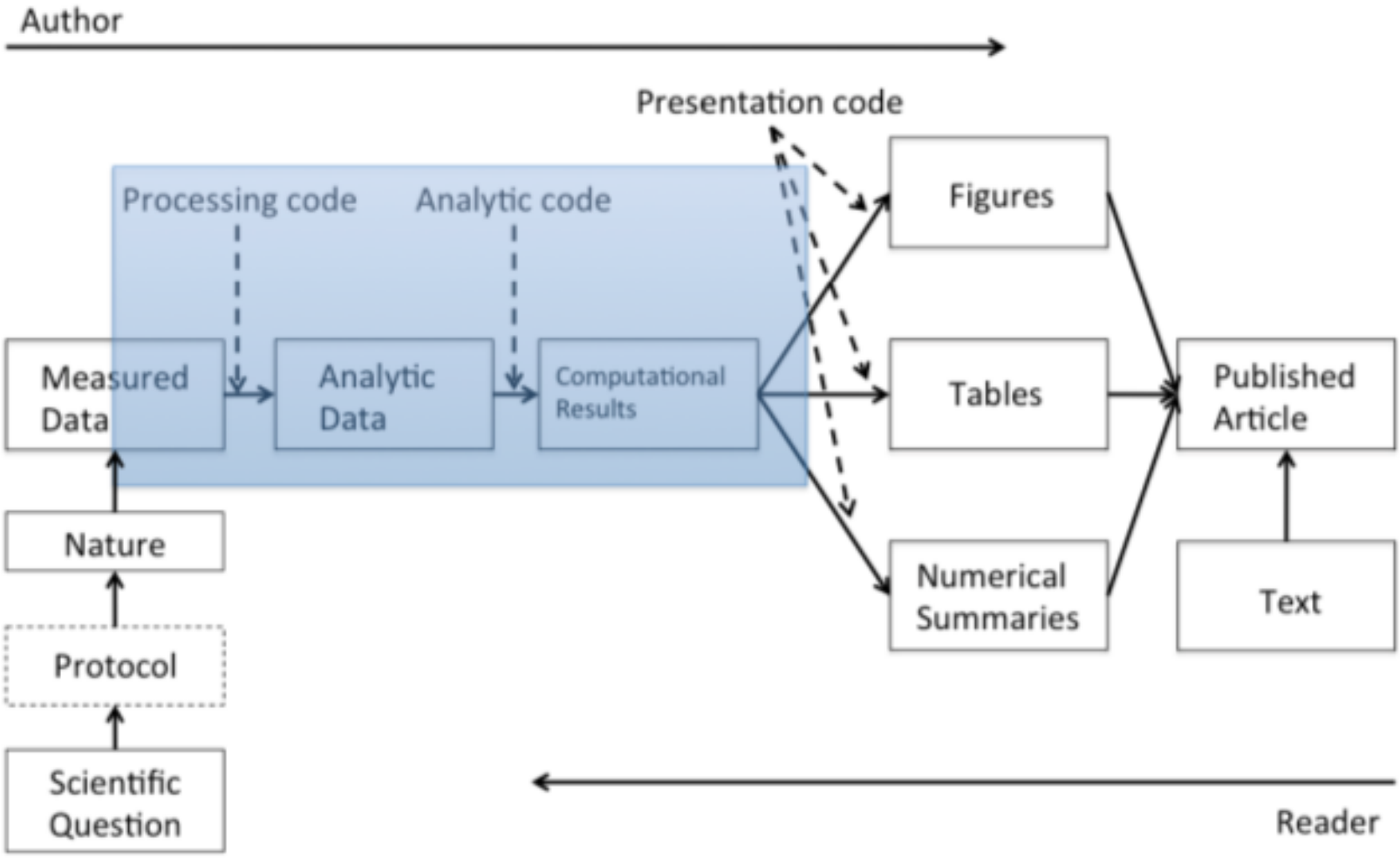


# What is Reproducible Research?

Reproducible research is the idea that data analyses, and more generally, scientific claims, are published with their data and software code so that others may verify the findings and build upon them.



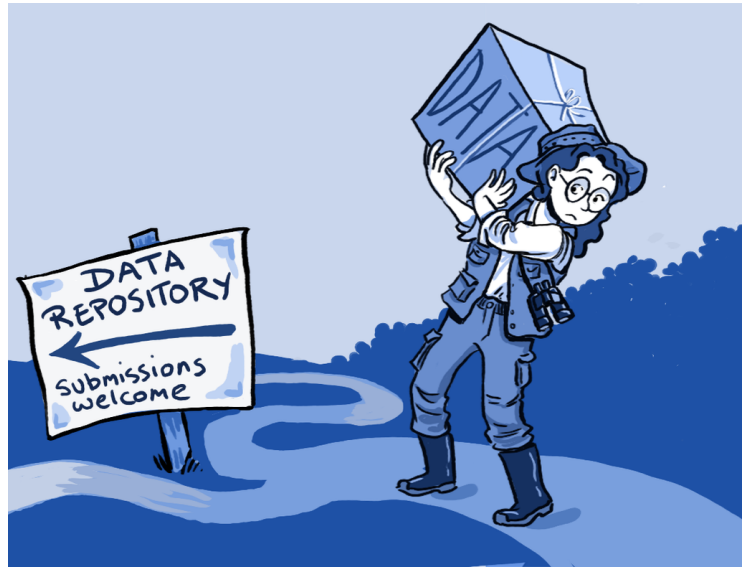
# What is Reproducible Research?



# What is Reproducible Research?

Related ideas / movements:

- open access
- open source
- open data
- literate programming




A.K.A : Open Science

"*It's a tragedy we had to add the word open to science.*"

# How did we end up here?

- "Science is in a crisis of (non) reproducibility."
- "I often found it difficult to replicate previous scientific results."
- "I was frustrated at my inability to identify the precise organisms, probes, antibodies and other scientific materials that underpinned genotype-phenotype assertions in the literature."
- "The lack of specificity in the literature was initially shocking to me"

*Source:* `peerj.com/about/author-interviews/`

# What could the reasons be?

- publication pressure, a feeling that there’s no time to "do it right"
- it is a fairly new phenomenon in science that experiments are run mainly / solely on computers: lack of accepted standards / routines for workflows
- some datasets are not free, or too big: not easy to handle without an expensive infrastructure
- many reserch papers are lacking details on purpose to make sure that a follow-up paper can NOT be done by someone else

# Why do we need Reproducible Research?

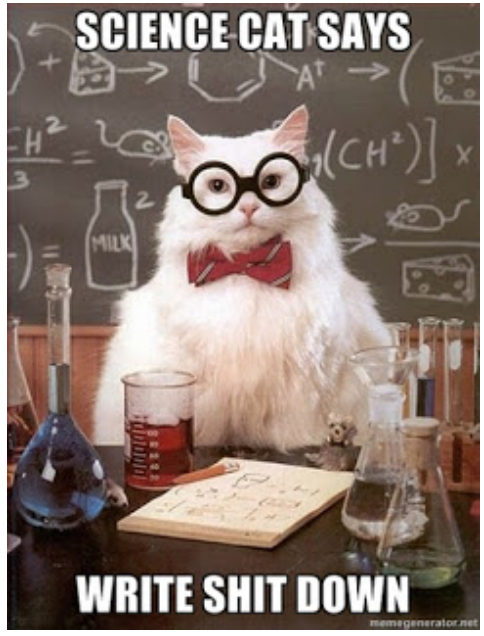
- to reduce the chances of embarrassing errors and faulty results
- to avoid multiplied efforts to reach the same results
- to save time (on the long run)
- to enable others to build upon it
- to increase public trust in science

# What has been done?

- Reproducibility manifesto
  `lorenabarba.com/gallery/reproducibility-pi-manifesto/`
- Coursera course on reproducible research
  `www.coursera.org/course/repdata`
- Publications about the issue (see later)
- More and more journals require publication of datasets and codes along with a paper

# What can WE do?

- at least write down everything you did (keep "lab notes")
- track & test & document your code
- publish in open access journals
- talk about the problem with other researchers
- take the "Reproducible Research" course on coursera :)



# What can WE do? - Manifesto 1

**The Reproducibility PI Manifesto**

1. I will teach my graduate students about reproducibility.
2. All our research code (and writing) is under version control.
3. We will always carry out verification and validation (V&V reports are posted to figshare)
4. For main results in a paper, we will share data, plotting script & figure under CC-BY

# The pledge - Manifesto 2

4. We will upload the preprint to arXiv at the time of submission of a paper.
5. We will release code at the time of submission of a paper.
6. We will add a "Reproducibility" declaration at the end of each paper.
7. I will keep an up-to-date web presence.

arXiv.org

# What are the obstacles / challenges?

Factors against reproducibility and open science:

- Laziness: it takes effort to make all data/results/code available
- Lack of convenient tools
- Lack of incentives
- Some are afraid of opening up their "lab notebooks" before everything is published, because someone might steal their ideas

# Summary

What is not reproducible is not science

# Related publications & sources I

- www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1003285
- www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0067111
- www.jove.com/blog/2012/05/03/studies-show-only-10-of-published-science-articles-are-reproducible-what-is-happening
- www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble
- phys.org/news/2013-09-science-crisis.html
- twitter.com/openscience/status/446942010554191872
- peerj.com/about/author-interviews/
- politicalsciencereplication.wordpress.com/2014/02/25/replication-workshop-what-frustrated-students-and-why-they-still-liked-the-course/
- www.wired.com/2014/07/incentivizing-peer-review-the-last-obstacle-for-open-access-science/

# Related publications & sources II

- yihui.name/en/2012/06/enjoyable-reproducible-research/
- yihui.name/slides/2012-knitr-RStudio.html#3.2
- biomickwatson.wordpress.com/2014/07/16/how-not-to-make-your-papers-replicable/
- kbroman.org/Tools4RR/assets/lectures/10_bigjobs_withnotes.pdf
- ivory.idyll.org/blog/ladder-of-academic-software-notsuck.html
- www.nature.com/nature/focus/reproducibility/
- ropensci.org/blog/2014/06/09/reproducibility/

Some more collected on the Twiki page:
info.ilab.sztaki.hu/twiki/bin/view/Main/ReproducibleResearch