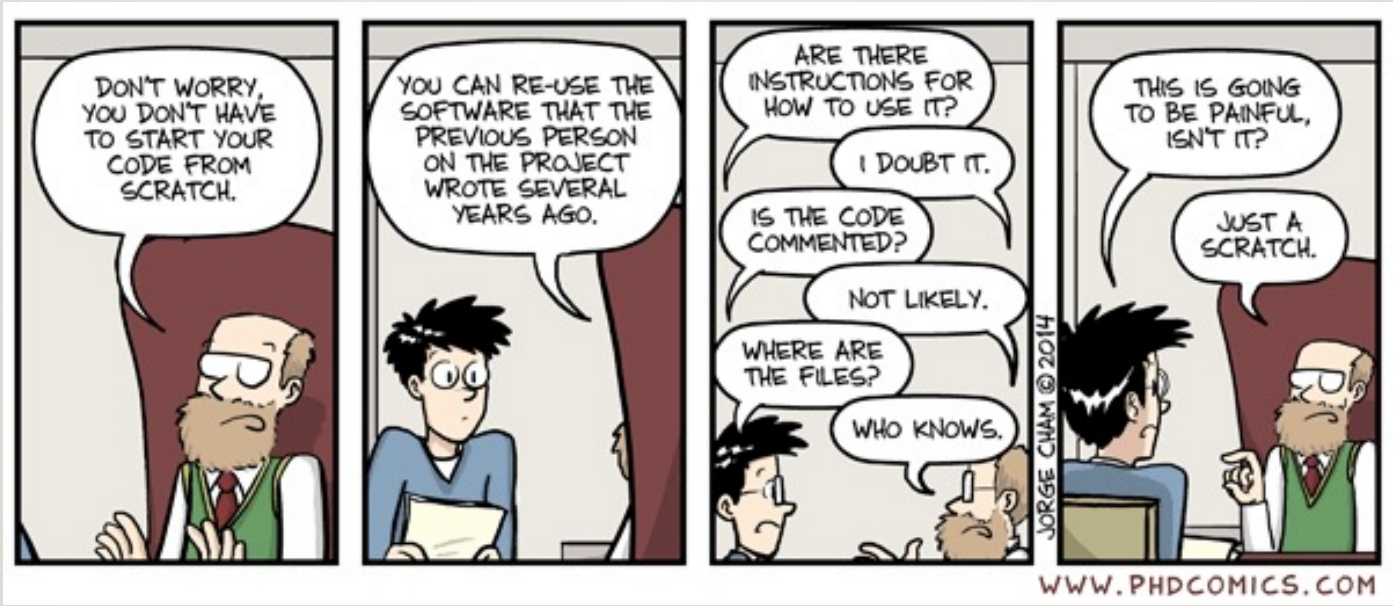
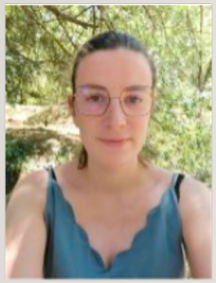
TOOLS FOR REPRODUCIBLE RESEARCH



The teachers



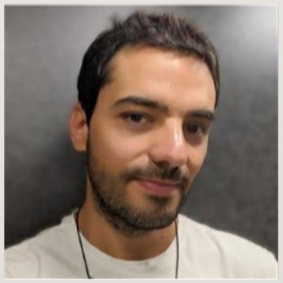
Aurore Comte

Jacques Dainat

Julie Orjuela

Gautier Sarah

Thomas Denecker



Nicolas Fernandez

Course content

- Good practices for working with data
- How to use the version control system **Git** to track changes to code
- How to use the package and environment manager **Conda**
- How to use the workflow managers **Snakemake** and **Nextflow**
- How to generate automated reports using **R Markdown**
- How to use **Jupyter** notebooks to document your analysis
- How to use **Docker** and **Singularity** to distribute containerized computational environments

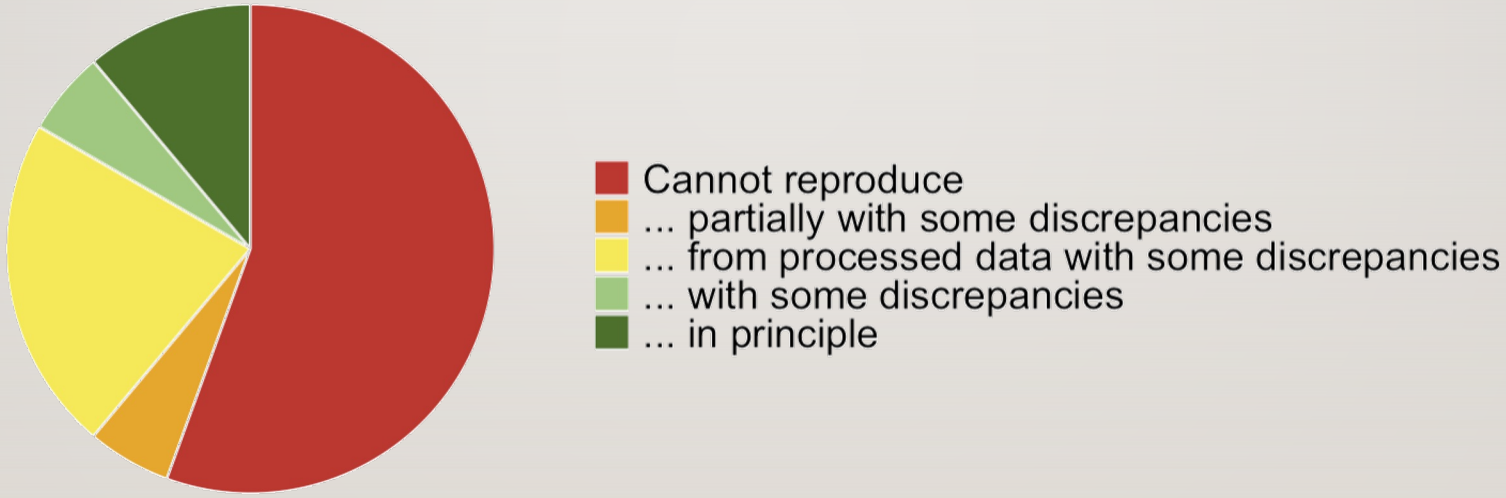
REPRODUCTIBILITY CRISIS

Crisis discovered middle 2000s
Getting bigger early 2010s
Many publications tackle the problem

REPRODUCTIBILITY CRISIS

Replication of data analyses in 18 articles on microarray-based gene expression profiling published in Nature Genetics in 2005–2006:

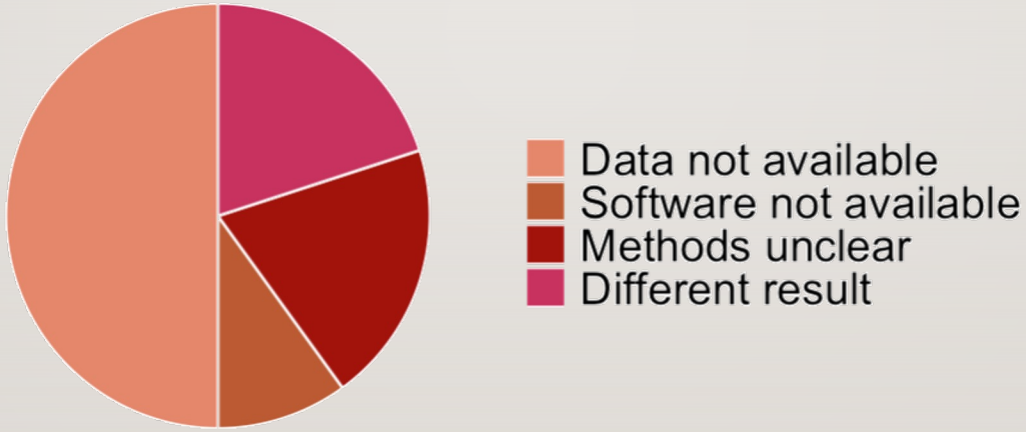
Adopted from Ioannidis *et al.* "Repeatability of published microarray gene expression analyses", Nature Genetics 41 (2009) doi:10.1038/ng.295



REPRODUCTIBILITY CRISIS

Replication of data analyses in 18 articles on microarray-based gene expression profiling published in Nature Genetics in 2005–2006:

Adopted from Ioannidis *et al.* "Repeatability of published microarray gene expression analyses", Nature Genetics 41 (2009) doi:10.1038/ng.295

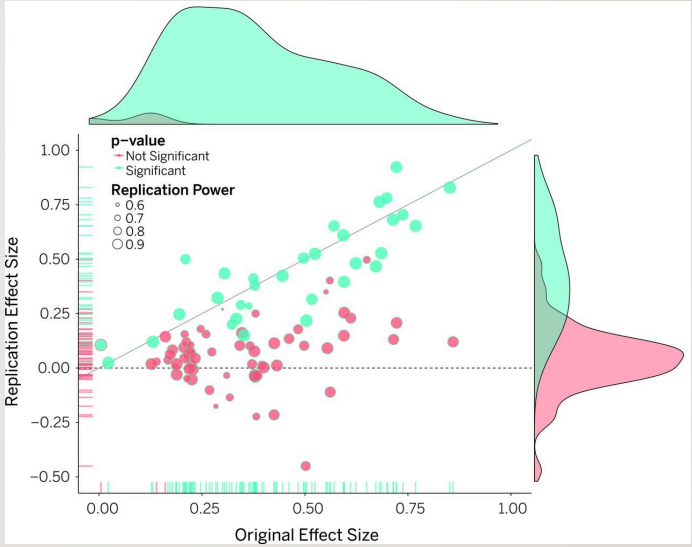


REPRODUCTIBILITY CRISIS

psychological science

The *Reproducibility project* set out to replicate 100 experiments published in high-impact psychology journals.*

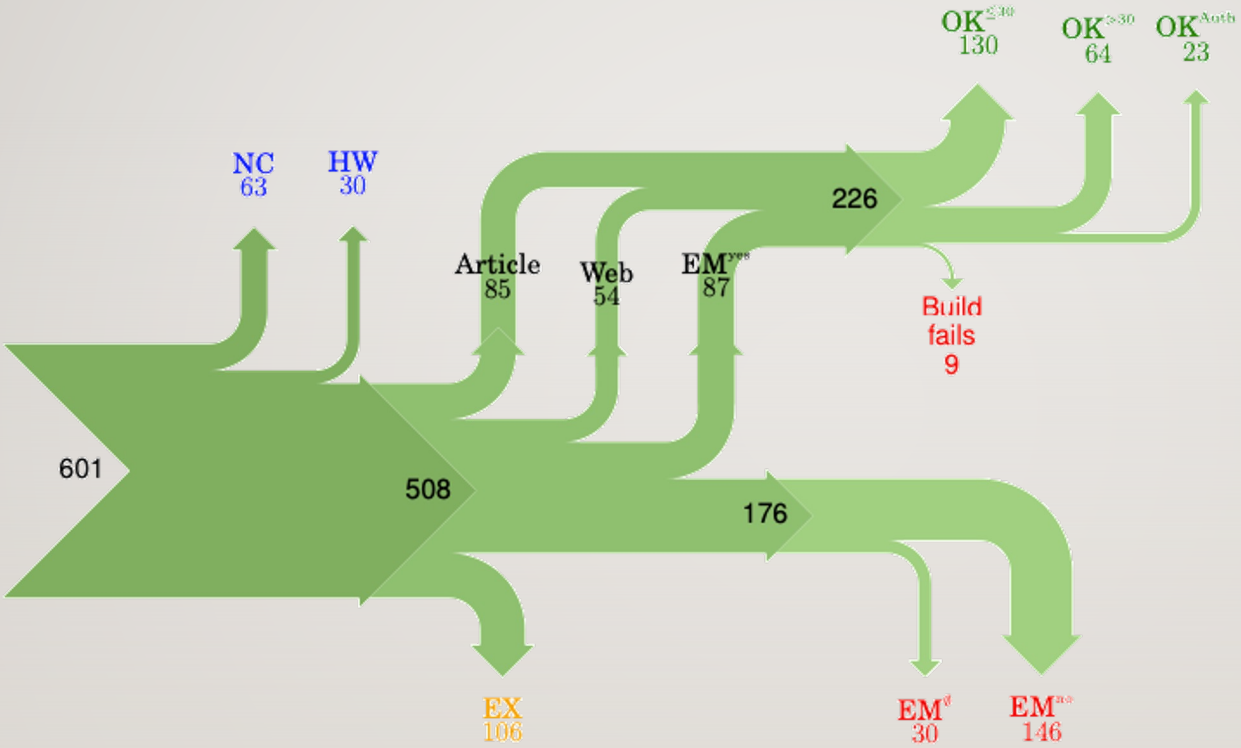
About one-half to two-thirds of the original findings could not be observed in the replication study.



* Open Science Collaboration. (2015). "Estimating the reproducibility of psychological science". Science. 349

REPRODUCTIBILITY CRISIS

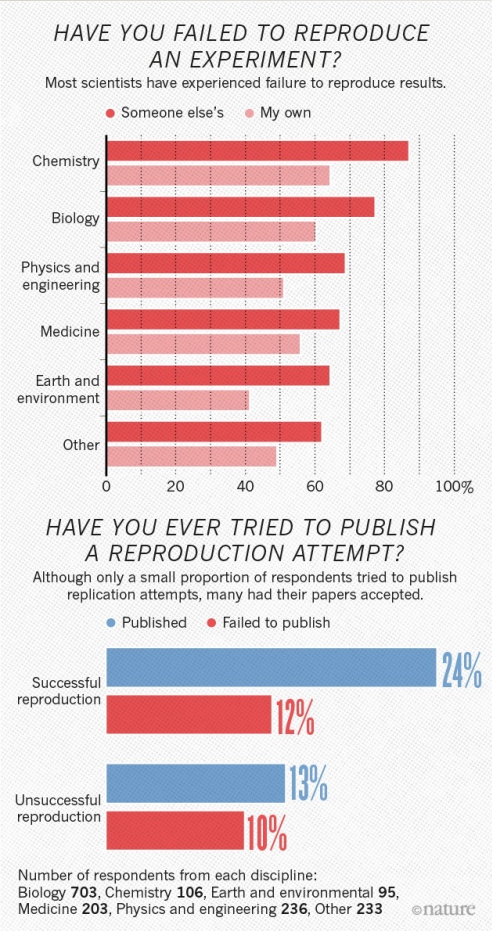
computer science



They succeeded to compile only **43%** of the codes

C. Colberg et. Al. 2015. Repeatability abd Benefaction in Computer Systems Research

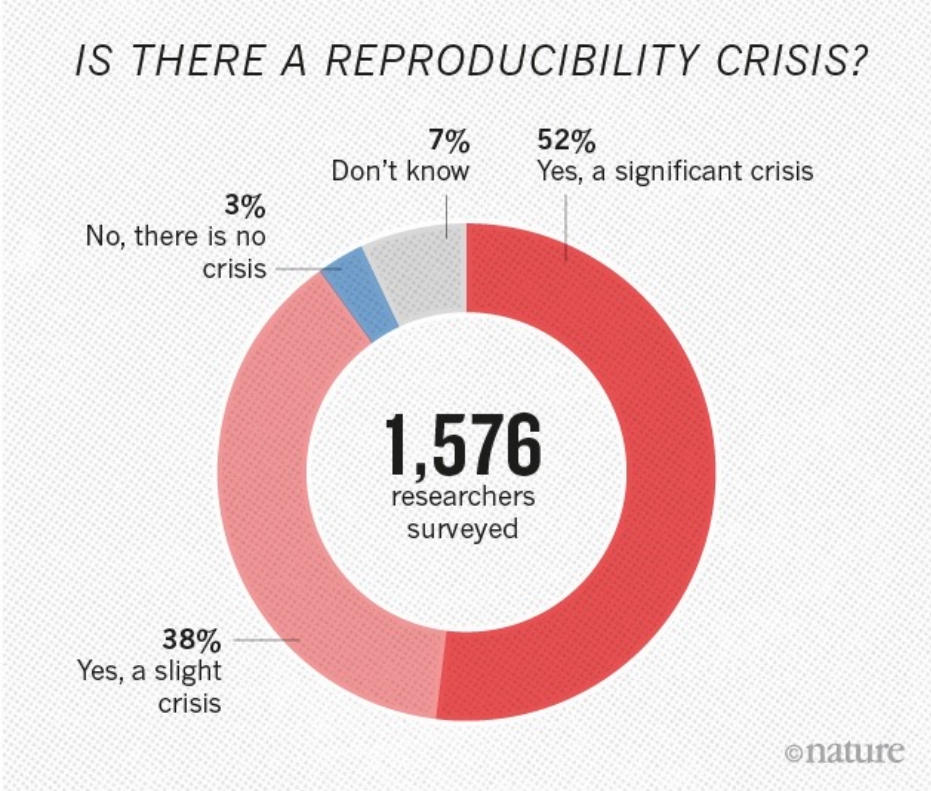
REPRODUCTIBILITY CRISIS



Nature 2016 (https://doi.org/10.1038/533452a)
1,500 scientists lift the lid on reproducibility

- A survey revealed that irreproducible experiments are a problem across all domains of science.
- More than 70% of researchers have tried and failed to reproduce another scientist's experiments
- More than half have failed to reproduce their own experiments.

REPRODUCTIBILITY CRISIS



Medicine is among the most affected research fields. A study in Nature found that 47 out of 53 medical research papers focused on cancer research were irreproducible*.

Common features were failure to show all the data and inappropriate use of statistical tests.

* Begley, C. G.; Ellis, L. M. (2012). "Drug development: Raise standards for preclinical cancer research". Nature. 483 (7391): 531–533

REPRODUCTIBILITY CRISIS

The results of only 26% out of 204 randomly selected papers in the journal *Science* could be reproduced.*

> *"Many journals are revising author guidelines to include data and code availability."*
>
> *"(...) an improvement over no policy, but currently insufficient for reproducibility."*

*Stodden et. al (2018). "An empirical analysis of journal policy effectiveness for computational reproducibility". PNAS. 115 (11): 2584-2589

REPRODUCTIBILITY CRISIS

Journal List > Sci Adv > v.7(21); 2021 May > PMC8139580

Science Advances | AAAS
SIGNIFICANT RESEARCH, GLOBAL IMPACT

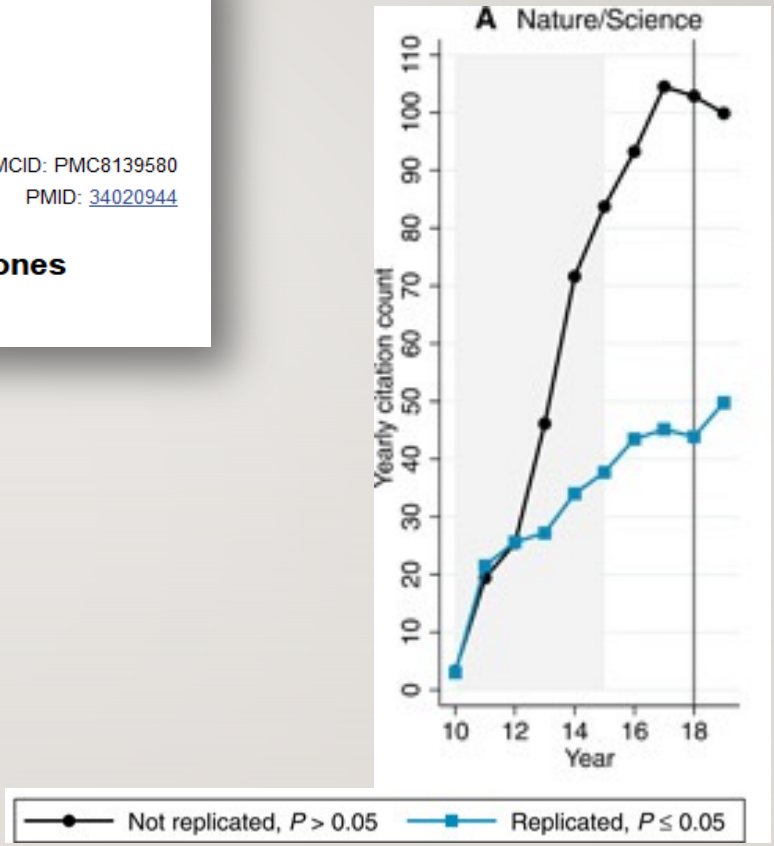
Sci Adv. 2021 May; 7(21): eabd1705.
Published online 2021 May 21. doi: 10.1126/sciadv.abd1705

PMCID: PMC8139580
PMID: 34020944

Nonreplicable publications are cited more than replicable ones

Marta Serra-Garcia*† and Uri Gneezy†

"Remarkably, only 12 percent of post-replication citations of non-replicable findings acknowledge the replication failure"



The average yearly citation count per year for studies that were not replicated.

REPRODUCTIBILITY CRISIS

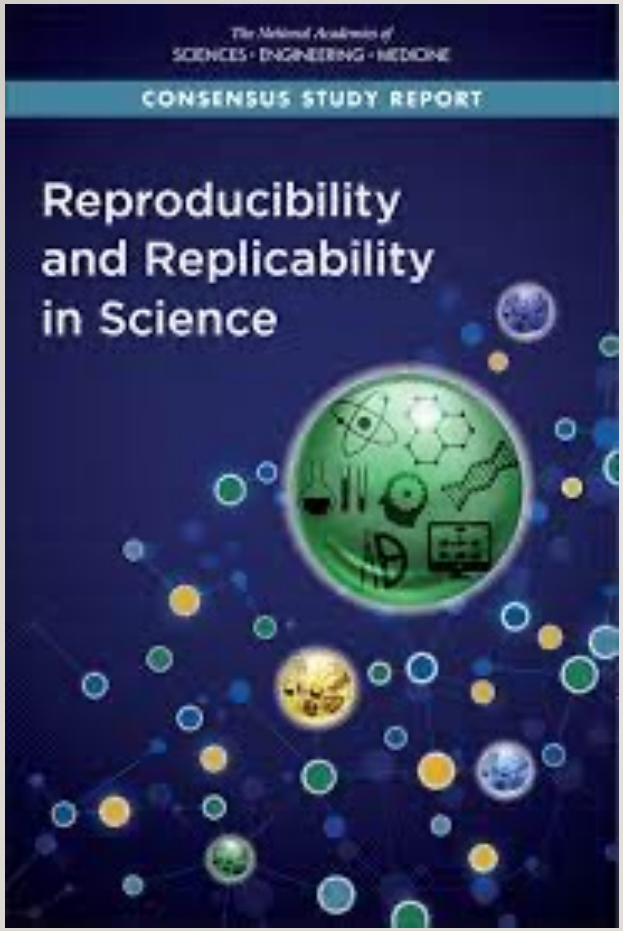
There are many so-called excuses not to work reproducibly:

> *"Thank you for your interest in our paper. For the [redacted] calculations I used my own code, and there is no public version of this code, which could be downloaded. Since this code is not very user-friendly and is under constant development I prefer not to share this code."*
>
> *"We do not typically share our internal data or code with people outside our collaboration."*
>
> *"When you approach a PI for the source codes and raw data, you better explain who you are, whom you work for, why you need the data and what you are going to do with it."*
>
> *"I have to say that this is a very unusual request without any explanation! Please ask your supervisor to send me an email with a detailed, and I mean detailed, explanation."*

Recurrent problems in code and environment

- Tools cannot be installed
  - Dependencies not available anymore
  - OS incompatible
- Tools/dependencies/libraries update break the code
  - tool arguments, function arguments, etc.
  - Python3 vs python2
- Result not-reproducible
  - Package versions
  - Scripts/commands lost or not saved
  - Code non reproductible (e.g seed)

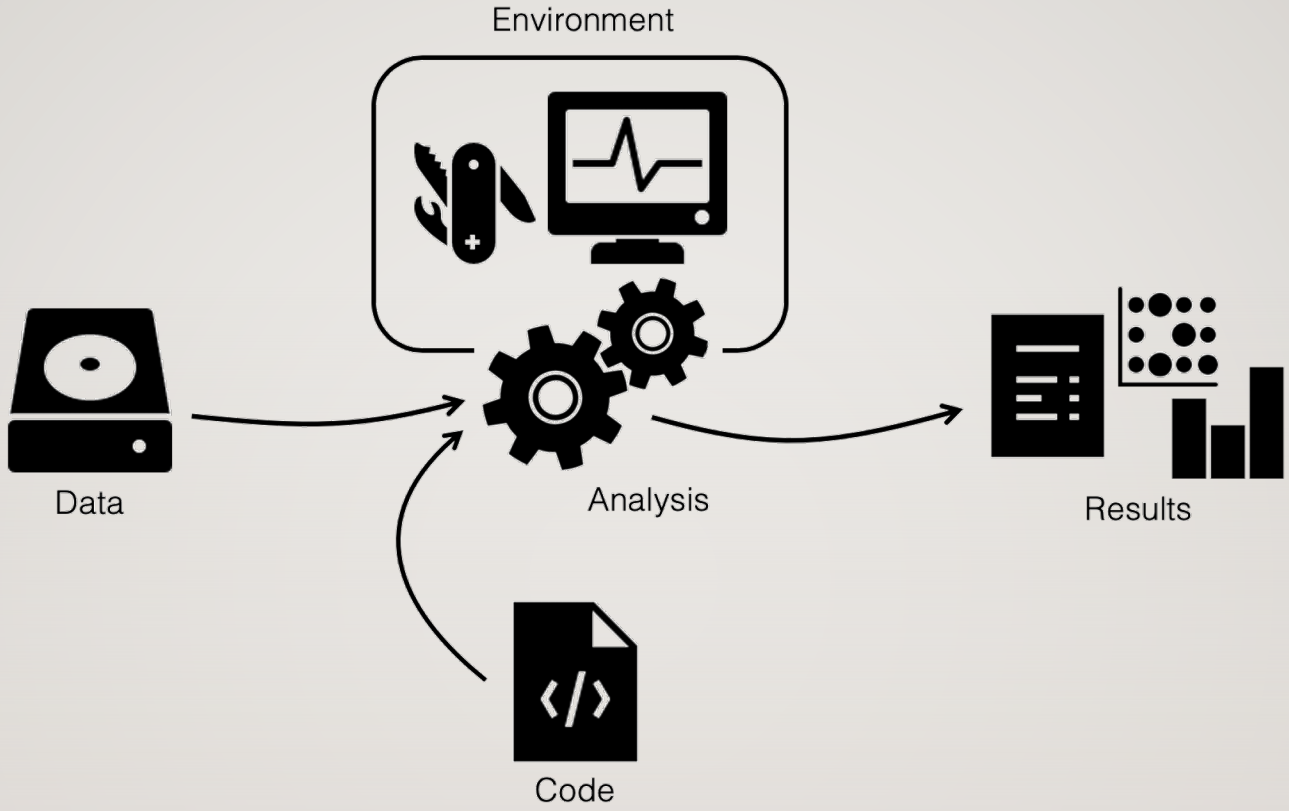
What does reproducible research mean?



National Academies of Sciences, Engineering, and Medicine. 2019. Reproducibility and Replicability in Science. https://doi.org/10.17226.25303

		Data	
		Same	Different
Code	Same	Reproducible	Replicable
	Different	Robust	Generalisable

What does reproducible research mean?



> *"Why call the course Reproducible Research, when it could just as well be called Research?"*
>
> *- Niclas Jareborg, NBIS data management expert*

How are you handling your data?

Decent

- Data available on request
- All metadata required for generating the results available

How are you handling your data?

Decent

- Data available on request
- All metadata required for generating the results available

Good

- Data deposited in public repositories
- Raw data available in unedited form
- If the raw data needed preprocessing, scripts were used rather than modifying it manually

How are you handling your data?

Decent

- Data available on request
- All metadata required for generating the results available

Good

- Data deposited in public repositories
- Raw data available in unedited form
- If the raw data needed preprocessing, scripts were used rather than modifying it manually

Great

- Section in the paper to aid in reproduction
- Used non-proprietary and machine-readable formats, e.g. .csv rather than .xls .

How are you handling your code?

Decent

- All code for generating results from processed data available on request

How are you handling your code?

Decent

- All code for generating results from processed data available on request

Good

- All code for generating results from raw data is available
- The code is publicly available with timestamps or tags

How are you handling your code?

Decent

- All code for generating results from processed data available on request

Good

- All code for generating results from raw data is available
- The code is publicly available with timestamps or tags

Great

- All code for generating results from publicly available raw data is available
- Code is documented and contains instructions for reproducing results
- Seeds were used and documented for heuristic methods

How are you handling your environment?

Decent

- Key programs used are mentioned in the methods section

How are you handling your environment?

Decent	• Key programs used are mentioned in the methods section
Good	• List of all programs used and their respective versions are available

How are you handling your environment?

Decent	• Key programs used are mentioned in the methods section
Good	• List of all programs used and their respective versions are available
Great	• Instructions for reproducing the environment publicly available

"What's in it for me?"

Before the project

- Improved structure and organization
- Forced to think about scope and limitations

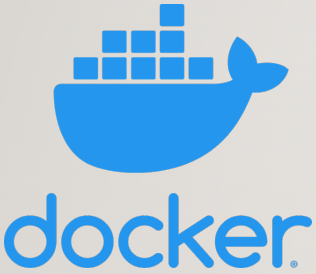
During the project

- Easier to re-run analyses and generate results after updating data, tools, parameters, etc.
- Closer interaction between collaborators
- Much of the manuscript "writes itself"

After the project

- Faster resumption of research by others (or, more likely, your future self), thereby increasing the impact of your work
- Increased visibility in the scientific community

Tools are here to help you



Tools are one thing, but you need to know how to you use them properly…

Questions?