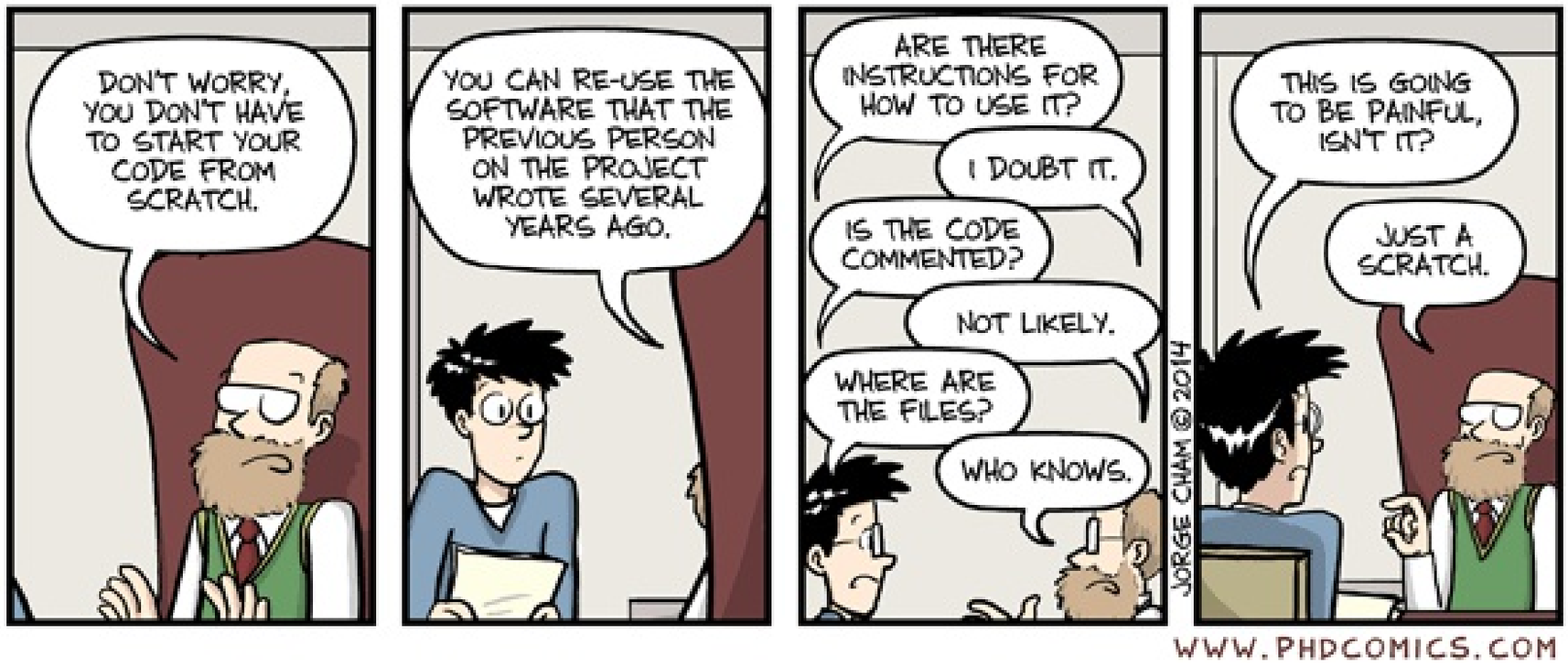
Tools for reproducible research



SciLifeLab

NBIS

The teachers

John Sundh

Verena Kutschera

Erik Fasterius

Tomas Larsson

SciLifeLab

NBIS

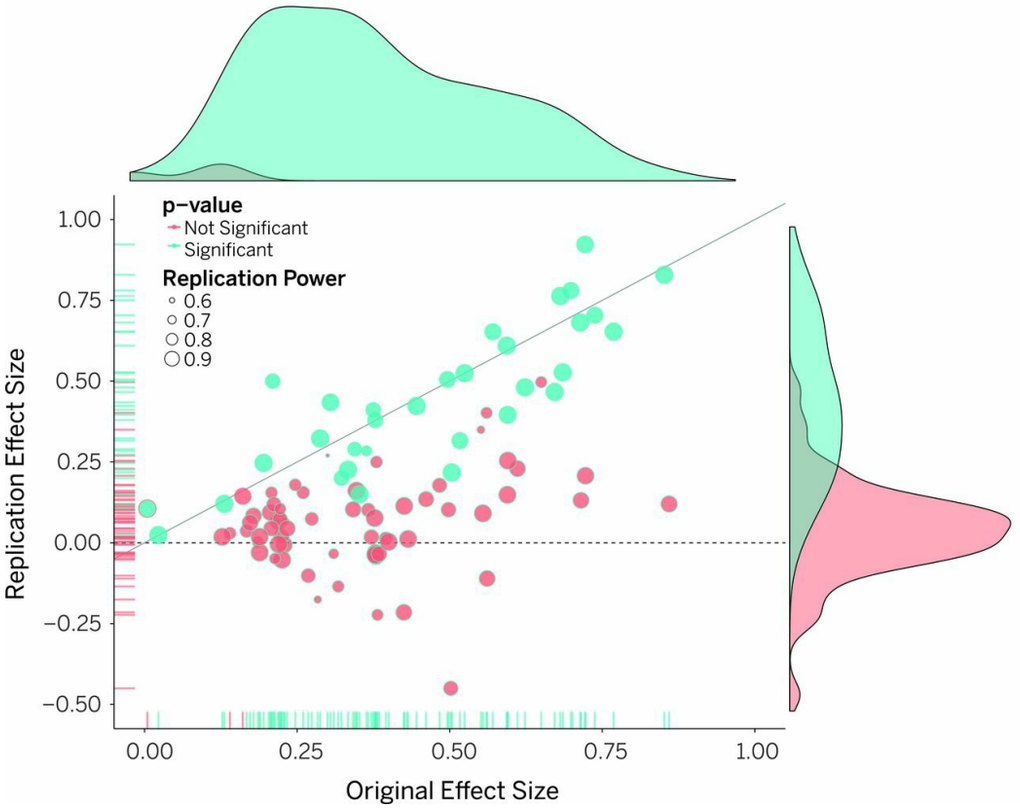
Course content

- 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 manager Snakemake
- How to use R Markdown to generate automated reports
- How to use Jupyter notebooks to document your analysis
- How to use Docker and Singularity to distribute containerized computational environments

Why all the talk about reproducible research?

The Reproducibility project set out to replicate 100 experiments published in high-impact psychology journals.[1]

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

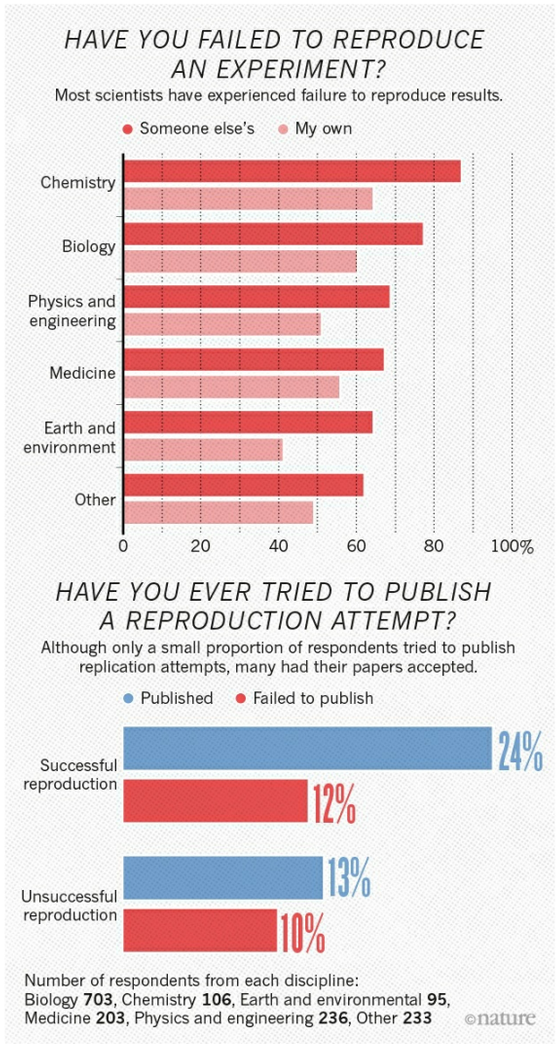


[1] "Estimating the reproducibility of psychological science". Science. 349

SciLifeLab

NBIS

Why all the talk about reproducible research?



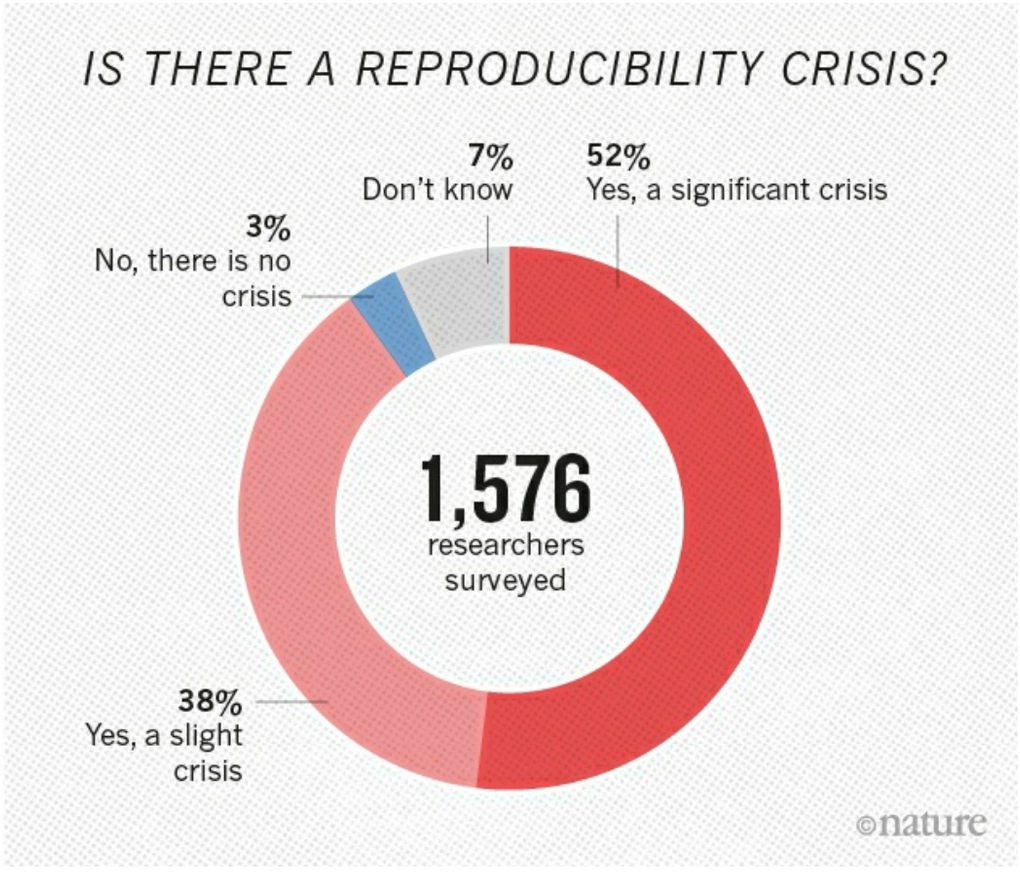
A survey in Nature revealed that irreproducible experiments are a problem across all domains of science.[1]

[1] "1,500 scientists lift the lid on reproducibility", Nature. 533: 452–454

SciLifeLab

NBIS

Why all the talk about reproducible research?



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.[2]

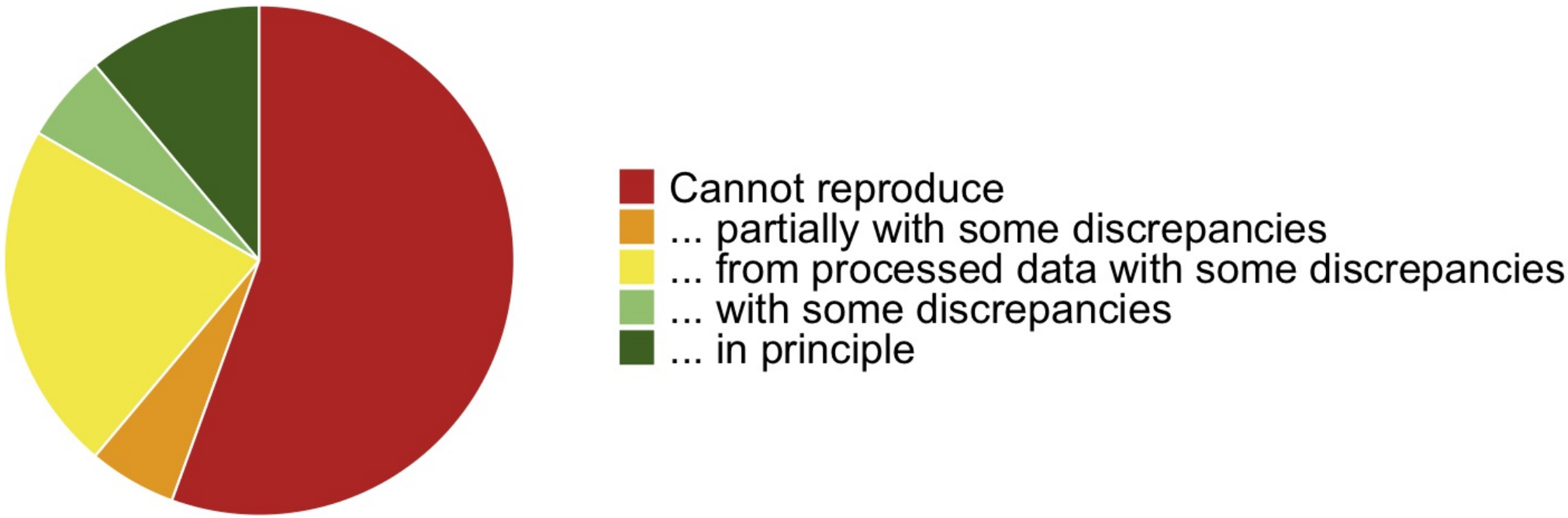
Common features were failure to show all the data and inappropriate use of statistical tests.

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

Why all the talk about reproducible research?

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



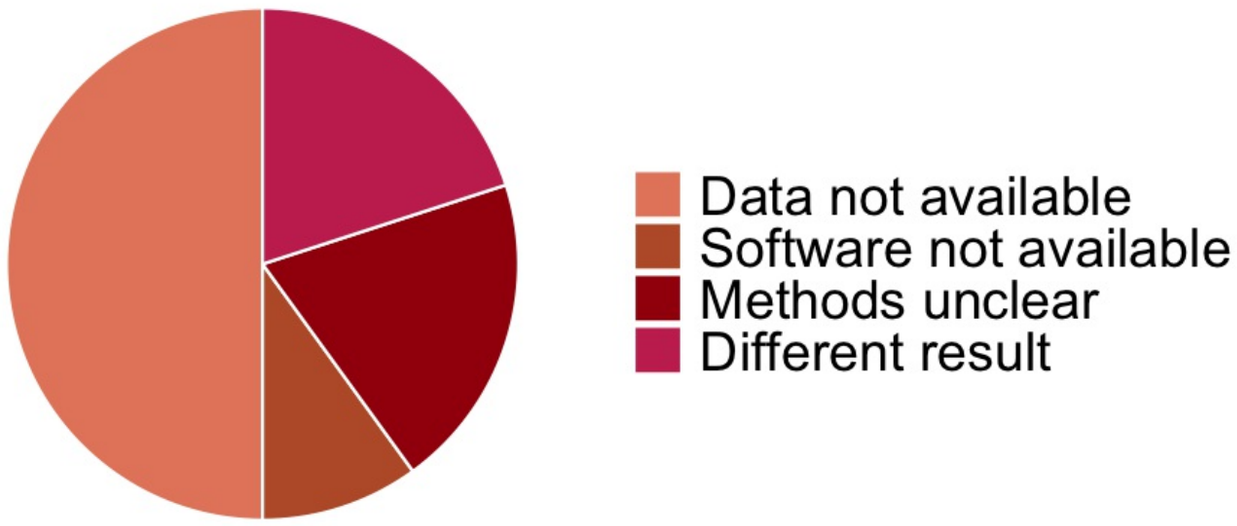
SciLifeLab

NBIS

Why all the talk about reproducible research?

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



Reproducibility is rarer than you think

The results of only 26% out of 204 randomly selected papers in the journal Science could be reproduced.[1]

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

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

SciLifeLab

NBIS

Reproducibility is rarer than you think

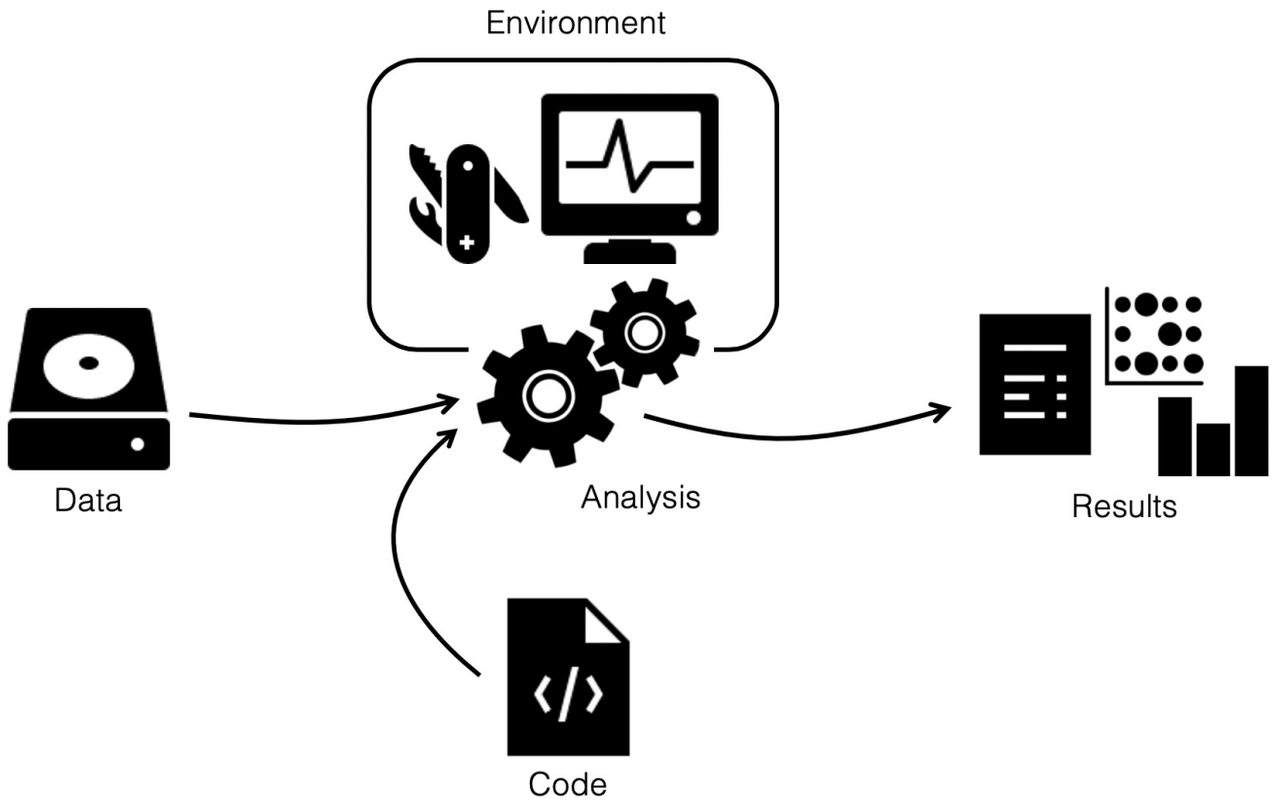
There are many so-called excuses not to work reproducibly:

> “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.”

What does reproducible research mean?

		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

SciLifeLab

NBIS

How are you handling your data?

Decent

Good

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

SciLifeLab

NBIS

How are you handling your data?

Decent

Good

Great

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

SciLifeLab

NBIS

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

Good

Great

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

Good

Great

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

SciLifeLab

NBIS

"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 rerun 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

Questions so far?