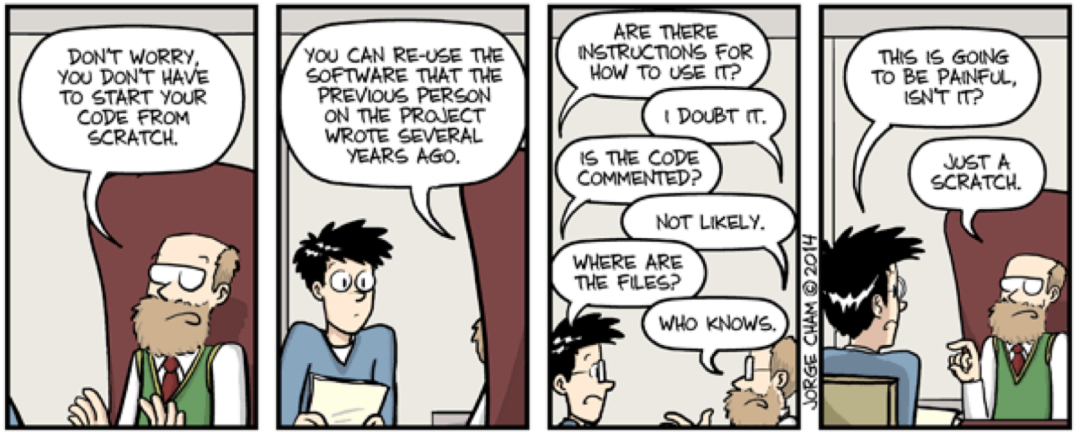
Tools for Reproducible Research

1 ECTS

Göteborg, November 28-29



"Piled Higher and Deeper" by Jorge Cham

NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

NBIS
NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

Teachers

Rasmus

Leif

Viktor

Course content:

- good practices for data analysis and management
- how to use the version control system git to track edits and collaborate on coding
- 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 ongoing analysis
- how to use Docker to distribute containerized computational environments

git
CONDA
Snakemake
R Markdown
jupyter
docker

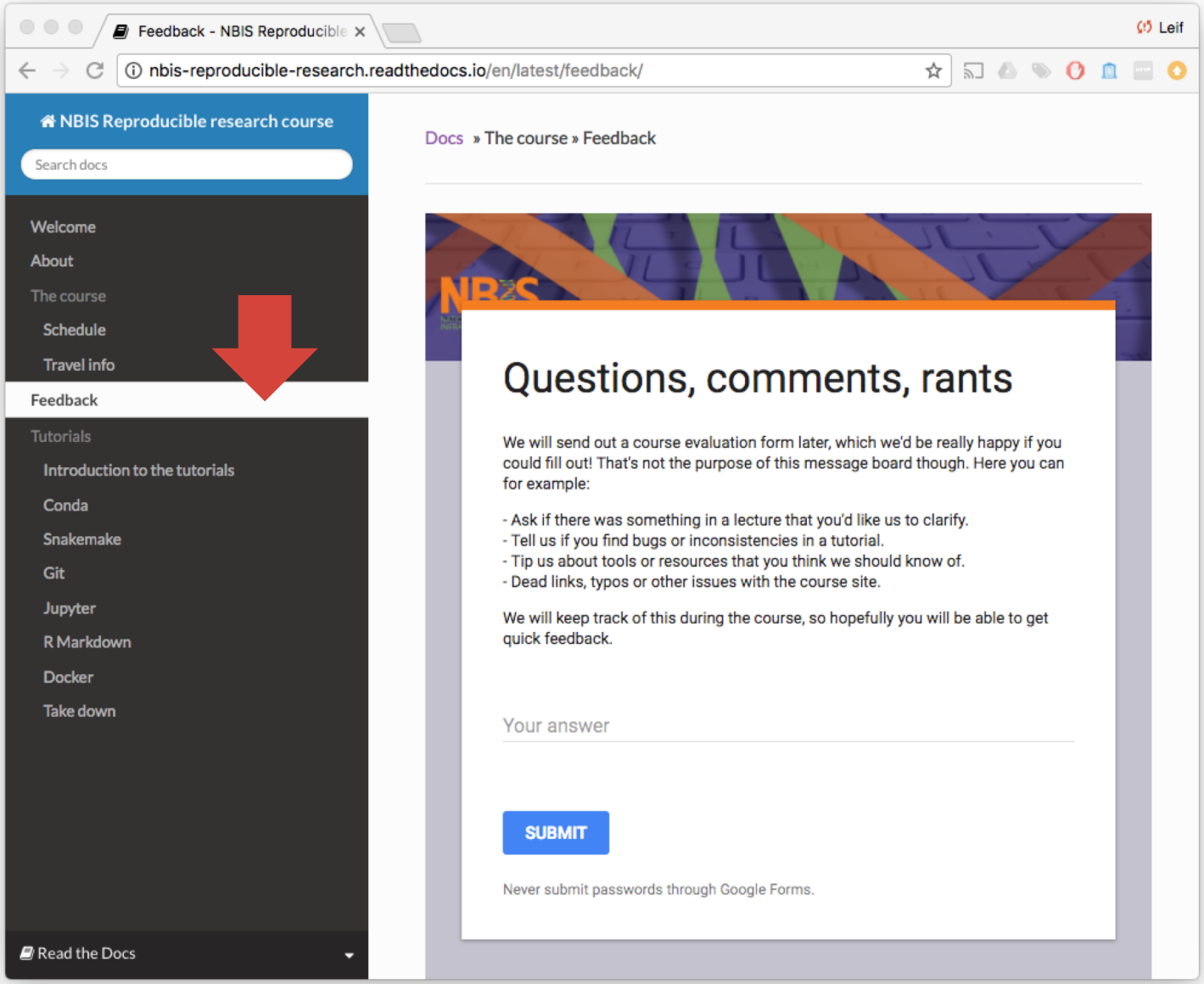
Schedule

Today

Time	Topic
09:00	Introduction to Reproducible Research
09:45	Data management and project organization
10:15	Fika break
10:45	Master your dependencies - environments and reproducibility *- Introduction to the package and environment manager Conda* *- Practical tutorial: Conda*
12:00	Lunch
13:00	Organize your analysis using workflow managers *- Introduction to Snakemake* *- Practical tutorial: Snakemake*
16:15	Wrap-up
16:30	Free time!
17:00	Joint departure for dinner
17:30	Dinner at Market

Tomorrow

Time	Topic
08:30	Distributing and version tracking your code *- Introduction to version control and git* *- Practical tutorial: Git* Computational notebooks *- Introduction to Jupyter* *- Practical tutorial: Jupyter*
12:00	Lunch
13:00	Reproducible reports *- Introduction to R Markdown* *- Practical tutorial: R Markdown* Containerization *- Introduction to containers* *- Practical tutorial: Docker*
16:00	Wrap-up
16:30	All done!



http://nbis-reproducible-research.readthedocs.io

Introduction to Reproducible Research

Why all the talk about

reproducible research?

RESEARCH ARTICLE

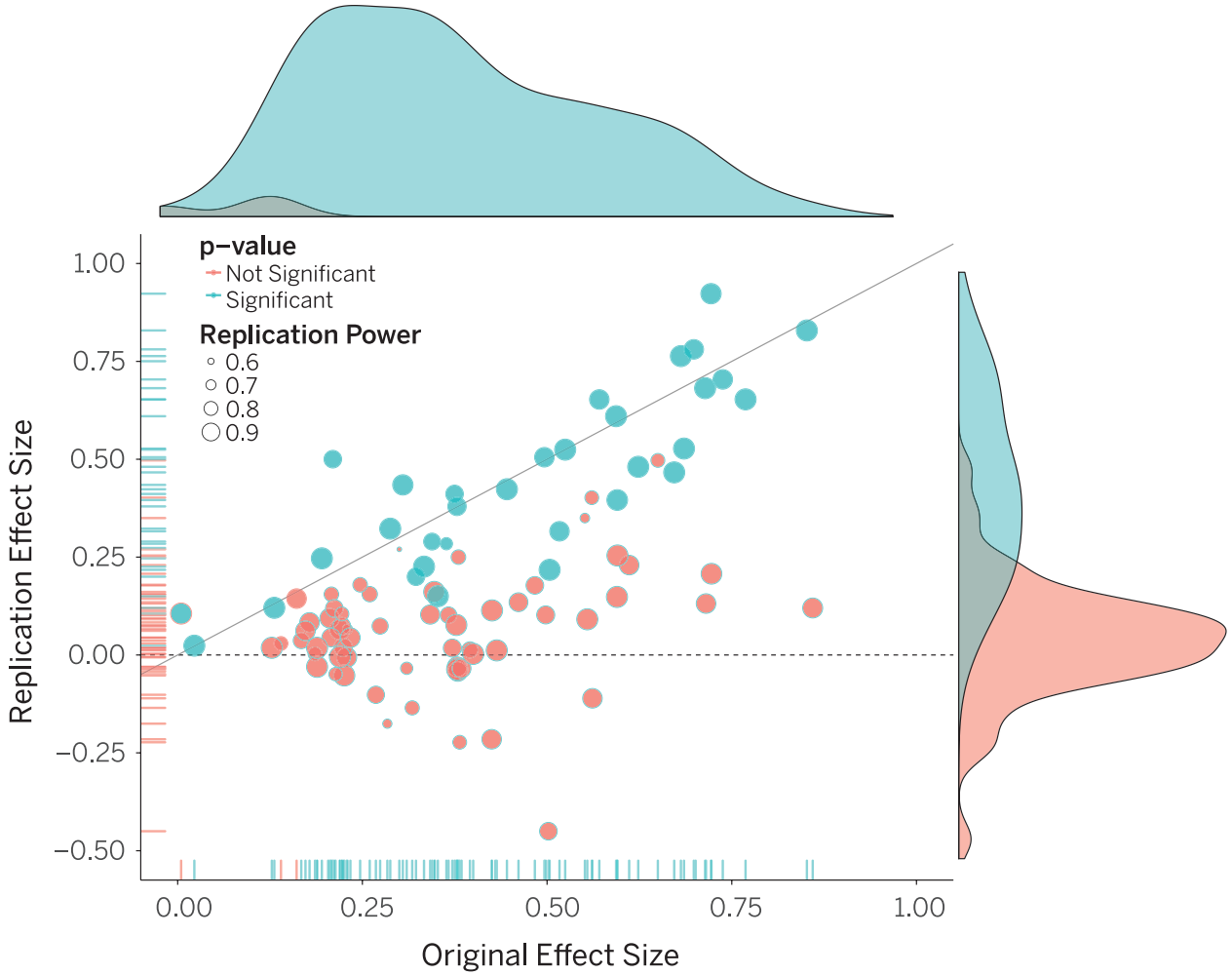
Estimating the reproducibility of psychological science

Open Science Collaboration*,†
+ See all authors and affiliations

Science 28 Aug 2015:
Vol. 349, Issue 6251, aac4716
DOI: 10.1126/science.aac4716

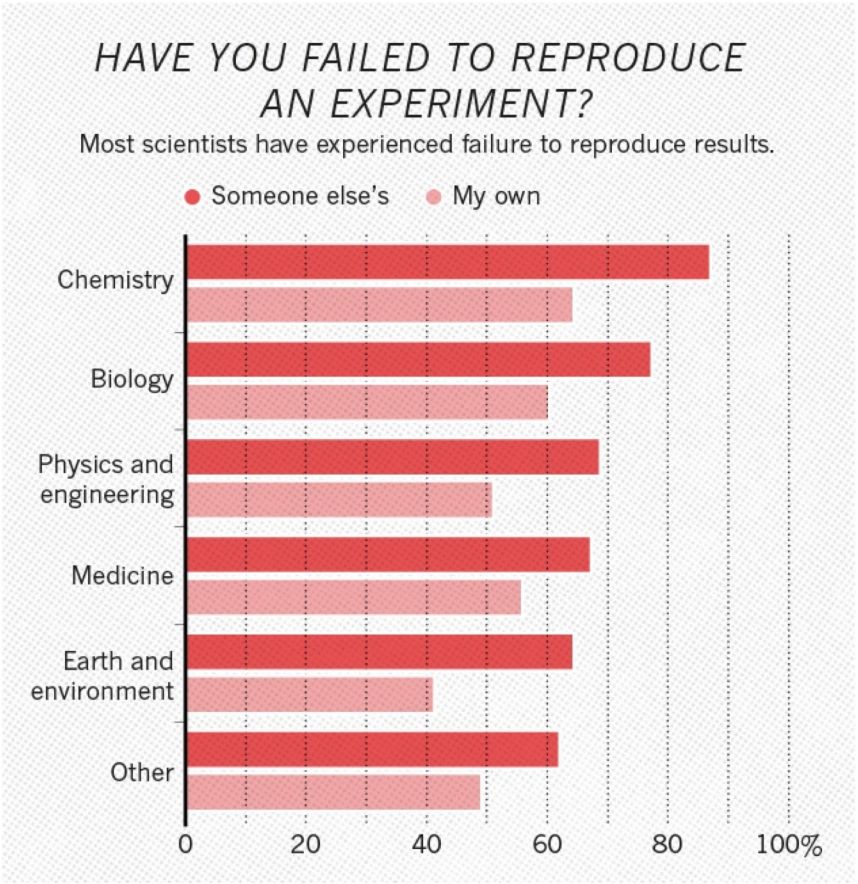
The *Reproducibility project* set out to replicate 100 experiments published in high-impact psychology journals.

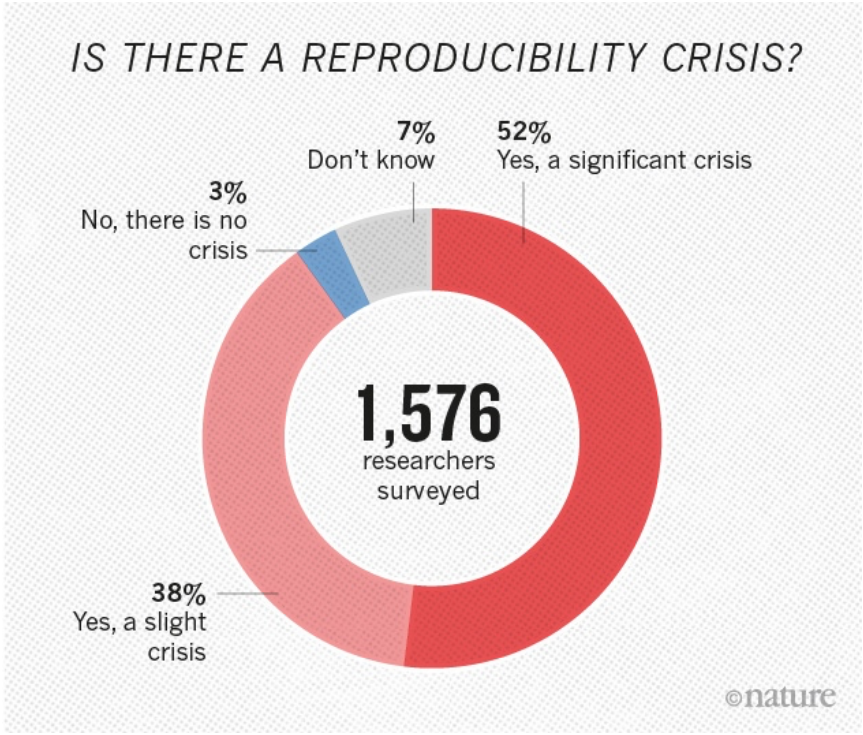
About one-half to two-thirds of the original findings could not be observed in the replication study.



Why all the talk about

reproducible research?





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

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].

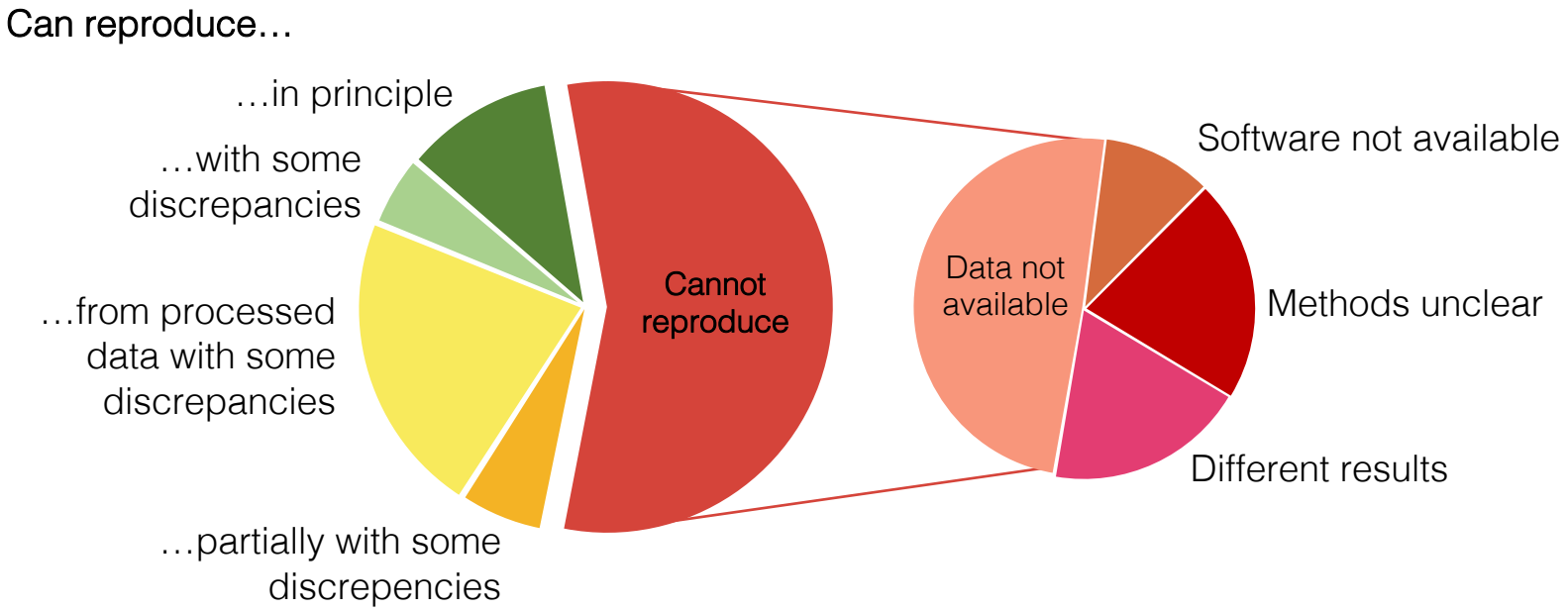
Common features were failure to show all the data and inappropriate use of statistical tests.

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



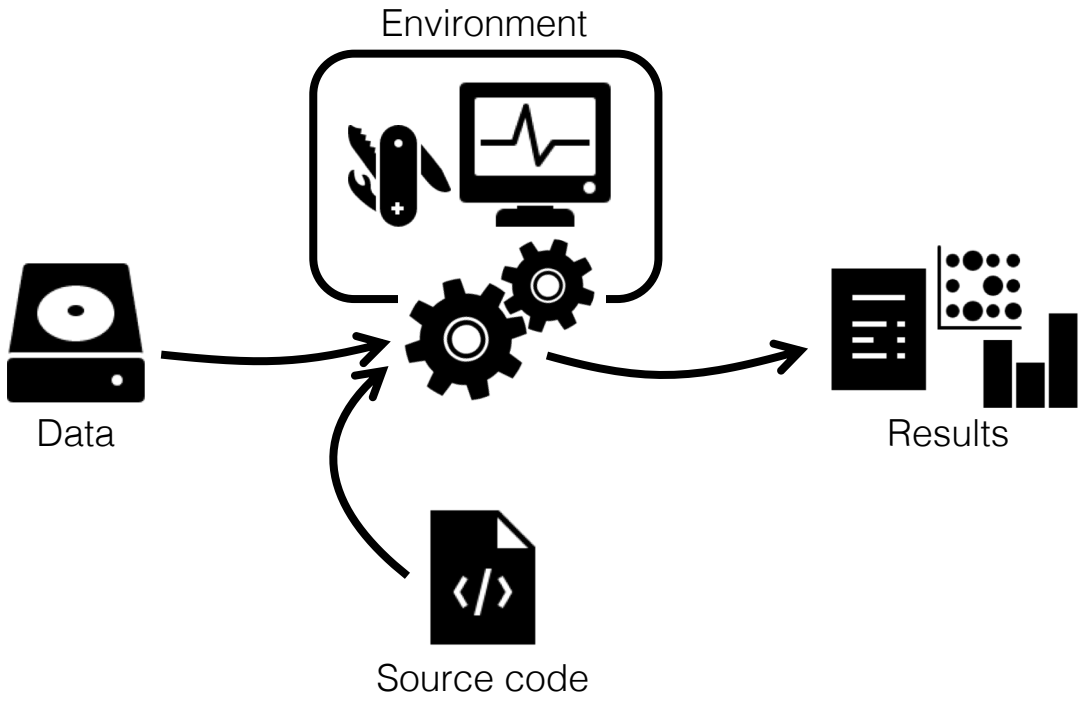
Summary of the efforts to replicate the published analyses.
Adopted from: Ioannidis et al. Repeatability of published microarray gene expression analyses.
Nature Genetics **41** (2009) doi:10.1038/ng.295

What do we mean with

reproducible research?

		Data: Same	Data: Different
Code	Same	Reproducible	Replicable
Code	Different	Robust	Generalisable

All parts of a bioinformatics analysis have to be reproducible:



”The foundations of knowledge should be constituted by experimentally produced facts, which can be made believable to a scientific community by their reproducibility."

- Robert Boyle, 1627-1691

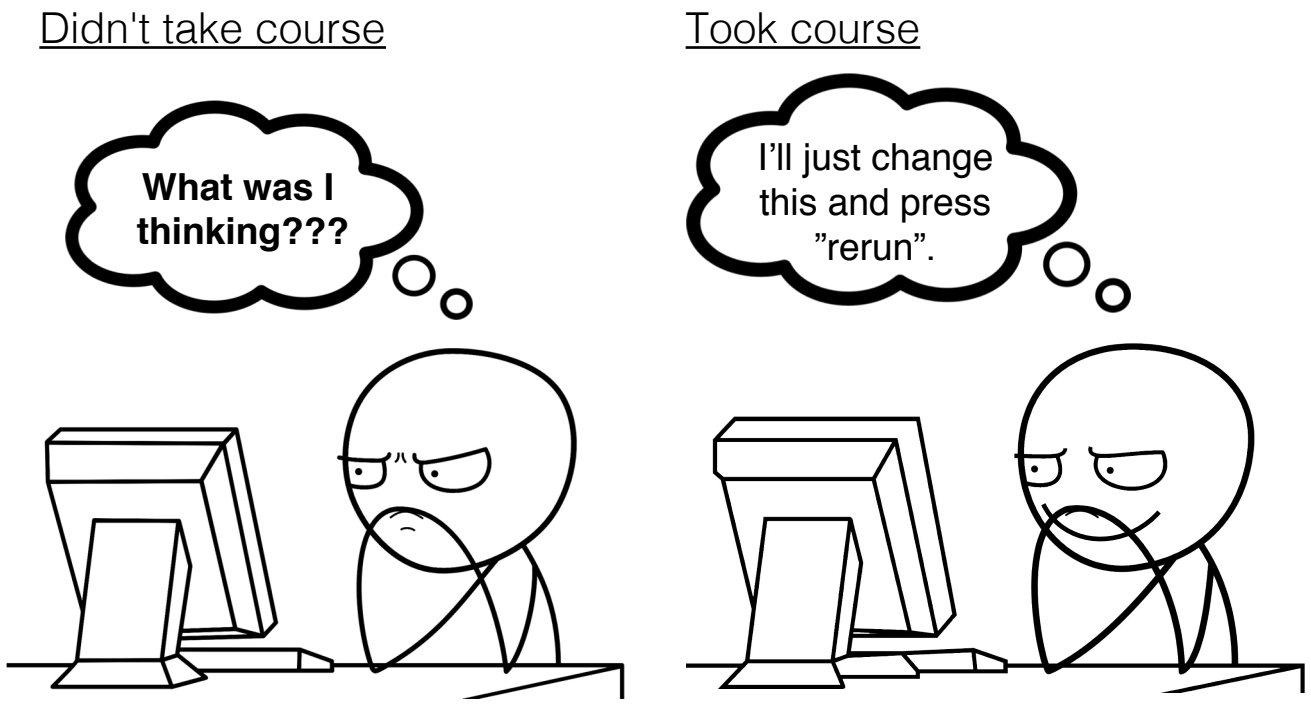
Where does your latest publication fit?

Decent	Getting there…	Well done!
• Data available on request. • All meta data 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.
• All code for generating results from processed data available on request.	• All code for generating results from raw data is available. • The code is publically available with timestamps/tags.	• All code for generating results from *publically available* raw data is available. • Code is documented and contains instructions for reproducing results. • Seeds were used and documented for heuristic methods.
• Key programs used are mentioned in the methods section.	• List of all programs used, and their respective versions, available.	• Instructions for reproducing the environment publically available.

Read more: Wilson et al. (2017) Good enough practices in scientific computing. PLoS Comput Biol 13(6)

What's in it for me?

One year in submission loop and reviewer comments are finally back…



"It takes some effort to organize your research to be reproducible. We found that although the effort seems to be directed to helping other people stand up on your shoulders, the principal beneficiary is generally the author herself. This is because time turns each one of us into another person, and by making effort to communicate with strangers, we help ourselves to communicate with our future selves."

Schwab et al. Making scientific computations reproducible. Computing in Science Engineering (2000).

Before project

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

During 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 project

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