ReproHacks:
Practicing reproducibility makes better

Anna Krystalli @annakrystalli

TUoS Open Research Conversation: Reproducibility and Preregistration
👋 Hello

me: Dr Anna Krystalli

- **Research Software Engineer**, *University of Sheffield*
  - twitter @annakrystalli
  - github @annakrystalli
  - email a.krystalli[at]sheffield.ac.uk

- **Editor rOpenSci**

- **Co-organiser**: Sheffield R Users group
Background
The paper is the advertisement

“an article about a computational result is advertising, not scholarship. The actual scholarship is the full software environment, code and data, that produced the result.”

John Claerbout paraphrased in Buckheit and Donoho (1995)

Why is our whole system geared towards reviewing, publishing, distributing, archiving the advertisement?

@annakrystalli
Progress: calls for reproducibility as minimum standard

Reproducibility has the potential to serve as a minimum standard for judging scientific claims when full independent replication of a study is not possible.

Reproducible Research in Computational Science ROGER D. PENG, SCIENCE 02 DEC 2011 : 1226-1227

@annakrystalli
Benefit #1
transparency as a means of verification

Benefit #2
transparency as a means of supercharging research cycle
So how are we doing?
If a paper claims to be reproducible but nobody checks it, is it really reproducible?
Practice

@annakrystalli
Reprohack

One day reproducibility hackathons

- How reproducible are papers?
- How can we provide a sandbox environment to practice reproducibility?

@annakrystalli
ReproHack History

OpenCon Satellite: Berlin, 2016

OpenCon Satellite: London, 2017

Inspired by Owen Petchey's Reproducible Research in Ecology, Evolution, Behaviour, and Environmental Studies course,

- Reproduce published results from raw data
- Over a few months and a number of sessions

ReproHack mission: Reproduce paper in a day from code and data

@annakrystalli
Software Sustainability Institute Fellowship 2019

@annakrystalli

Me: 👋
Let's run some ReproHacks!

C"Mcr19
CarpentryConnect Manchester 2019
ReproHacks since the Fellowship

- Leiden ReproHack
- N8 CIR Northern Tour ReproHack Series (x5)
- N8 CIR Remote ReproHack
- LatinR ReproHack
- UCL ReproHack for Open Access week

@annakrystalli
Reprohack Core Team

@annakrystalli
How does it work?
Call for papers

✨ Do you champion #reproducible #research?
✨ Do you have a reproducible paper with open code and data?

The @SoftwareSaved #ReproHack series needs you!

🚀 Help others learn & engage with your work by submitting it to our 1-day Reproducibility hackathons! https://t.co/PssdXqw18Z

– annakrystalli (@annakrystalli) June 12, 2019

@annakrystalli
On the day

- Select paper and form groups
- Work with materials and reproduce
- Discuss
- Feed back to authors
Tips for Reproducing & Reviewing

@annakrystalli
Selecting Papers

- Information submitted by authors:
  - Languages / tools used
  - Why you should attempt the paper.
- No. attempts No. times reproduction has been attempted
- Mean Repro Score Mean reproducibility score (out of 10)
  - lower == harder!

@annakrystalli
Review as an auditor

@annakrystalli
Access

- How easy was it to gain access to the materials?
- Did you manage to download all the files you needed?

Installation

- How easy / automated was installation?
- Did you have any problems?
- How did you solve them?
Data

- Were data clearly separated from code and other items?
- Were large data files deposited in a trustworthy data repository and referred to using a persistent identifier?
- Were data documented ...somehow...

Documentation

Was there adequate documentation describing:

- how to install necessary software including non-standard dependencies?
- how to use materials to reproduce the paper?
- how to cite the materials, ideally in a form that can be copy and pasted?

@annakrystalli
Analysis

- Were you able to fully reproduce the paper? ✅
- How automated was the process of reproducing the paper?
- How easy was it to link analysis code to:
  - the plots it generates
  - sections in the manuscript in which it is described and results reported

If the analysis was not fully reproducible 🚫

- Were there missing dependencies?
- Was the computational environment not adequately described/captured?
- Was there bugs in the code?
- Did code run but results (e.g., model outputs, tables, figures) differ to those published? By how much?
Review as a user 🎮

New User

Invested User

@annakrystalli
Feedback as a community member

- Acknowledge author effort
- Give feedback in good faith
- Focus on community benefits and system level solutions

Help build convention on what form a Reproducible paper should take and how we should be able to use it

@annakrystalli
What did we learn?
N8 CIR ReproHack Series Stats

- 38 papers submitted so far
- Total of ~ 70 participants
- 39 completed reviews over 27 papers

@annakrystalli
Review Scores

@annakrystalli
What would improve reproducibility?

- Better documentation
- More complete description / capture of computational environment

What was their favourite aspect of reproducible materials?

- Literate programming

@annakrystalli
Opportunity for peer skill sharing

- CCMcr: Contributing to open source
- Leiden: Synching GitHub repositories with Zenodo
- Remote Reprohack: Docker school

Fit for purpose

On the way home, @df3n5 said quite rightly, if all [code-producing/data-analysing] researchers would take part in at least one @ReproHack, the code reproducibility and quality of documentation would generally soar!

— Durham University Advanced Research Computing (@ARC_DU)
January 22, 2020

@annakrystalli
ReproHacks are fun

@annakrystalli
On the future of Reviewing

@annakrystalli
REVIEWER

I found and corrected some typos and a minor error in the code. Results are updated.

I reanalysed the parameter X and found it significantly affected the results.

I think you should perform complicated analysis.

AUTHOR

😊 Thanks!

WOW! That's an interesting finding. It will probably require a significant rewrite. Would you like to collaborate on an updated manuscript?

We feel that's beyond the scope of this paper. The code and data are freely available. Let us know if you wish to collaborate on a new paper.

@annakrystalli
On the scope of reproducibility

- Reproducibility *ad infinitum*
  - ✗ UNREALISTIC

@annakrystalli
On the scope of reproducibility

- Reproducibility *ad infinitum*
  - ❌ UNREALISTIC

- Reproducibility for 2-3 years post-publication
  - ✔️ MORE REALISTIC
    - Checked as part of publication process, e.g. CODE CHECK
      - https://codecheck.org.uk/

@annakrystalli
On the scope of reusability

Openness can help:

- surface useful parts of code.
- facilitate user feedback and contribution

MAINTENANCE?!

@annakrystalli
...in the meantime

take any opportunity to practice!
Are the participants geographically located in the same place?

**YES**

**Event ReproHack**

👩‍🎓 Conference  🏛️ University

- Team with people of different backgrounds.
- Decide which paper reproduce from a variety of options.
- Networking.

**Research Group ReproHack**

👩‍🔬 Team

- Let your team reproduce you article before is submitted.
- Reproduce papers related to your research topic
- Improve the capabilities of your team in scientific reproducibility.

**NO**

**Remote ReproHack**

@annakrystalli

- Participants can join as a research group or work together in a particular paper selecting different breaking rooms.
- It allows the presence of scientists around the world.
Ways to participate

Propose a paper
You've put a lot of effort into making your work reproducible. Now let people learn from and engage with it!

Benefits to authors:
- Feedback on the reproducibility of your work.
- Appreciation for your efforts in making your work reproducible.
- Opportunity to engage others with your research.

Submit paper!

Reproduce
Join a ReproHack and get working with other people's material!

Benefits to participants:
- Practical experience in reproducibility with real published materials
- Opportunity to explore different tools and strategies.
- Opportunity to for meaningful contribution.
- Inspiration to work more openly.

Join an event!

Organise an event
Help create a practical learning space

Benefits to community:
- Help build capacity in reproducibility throughout the research community.
- Highlight community value of reproducibility beyond validation of results.
- Help community evaluate how successful current practices are and for what purpose.
- Help identify what works and where the most pressing weaknesses in our approaches are.

Submit an event!
Interested in ReproHacking?

reprohack/reprohack-hq GH repository

Chat to us:

Host your own event!

Submit your own papers!

@annakrystalli
👋 Thanks for 🎉

❓

@annakrystalli
Resources

- **The Turing Way**: a lightly opinionated guide to reproducible data science.

- **Statistical Analyses and Reproducible Research**: Gentleman and Temple Lang's introduction of the concept of Research Compendia

- **Packaging data analytical work reproducibly using R (and friends)**: how researchers can improve the reproducibility of their work using research compendia based on R packages and related tools

- **How to Read a Research Compendium**: Introduction to existing conventions for research compendia and suggestions on how to utilise their shared properties in a structured reading process.

- **Reproducible Research in R with rrtools**: Workshop: Create a research compendium around materials associated with a published paper (text, data and code) using rrtools.
  - **Example Compendium**: Demo Research compendium.

@annakrystalli
Acknowledgements

Images throughout the slides watermarked with Scriberia were created by Scriberia for The Turing Way community and is used under a CC-BY licence.


Photo on slide #24 by Annie Spratt on Unsplash

Photo on slide #25 Sharon McCutcheon on Unsplash

@annakrystalli