ReproHacks:

Practicing reproducibility makes better

Anna Krystalli **@annakrystalli**

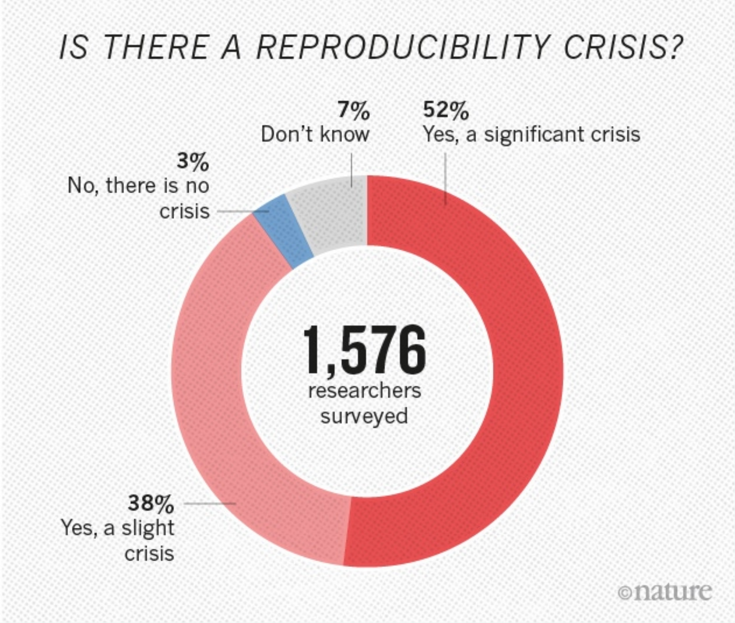
TUoS Open Research Conversation: Reproducibility and Preregistration

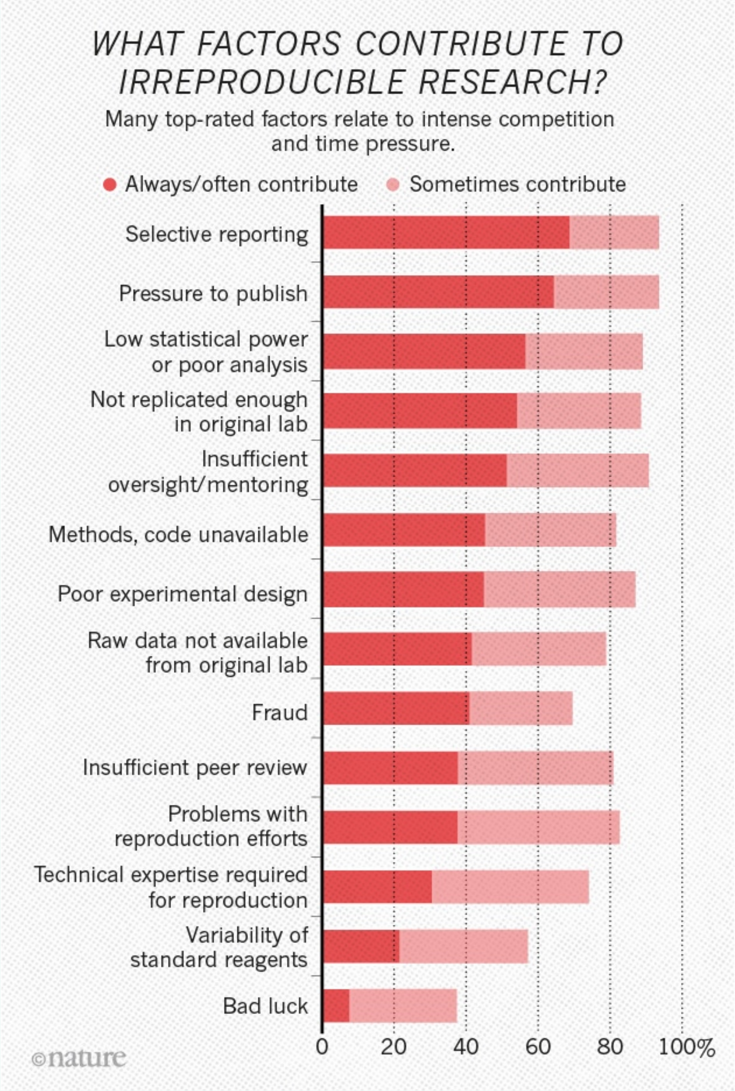
👋 Hello

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





@annakrystalli

The paper is the advertisement

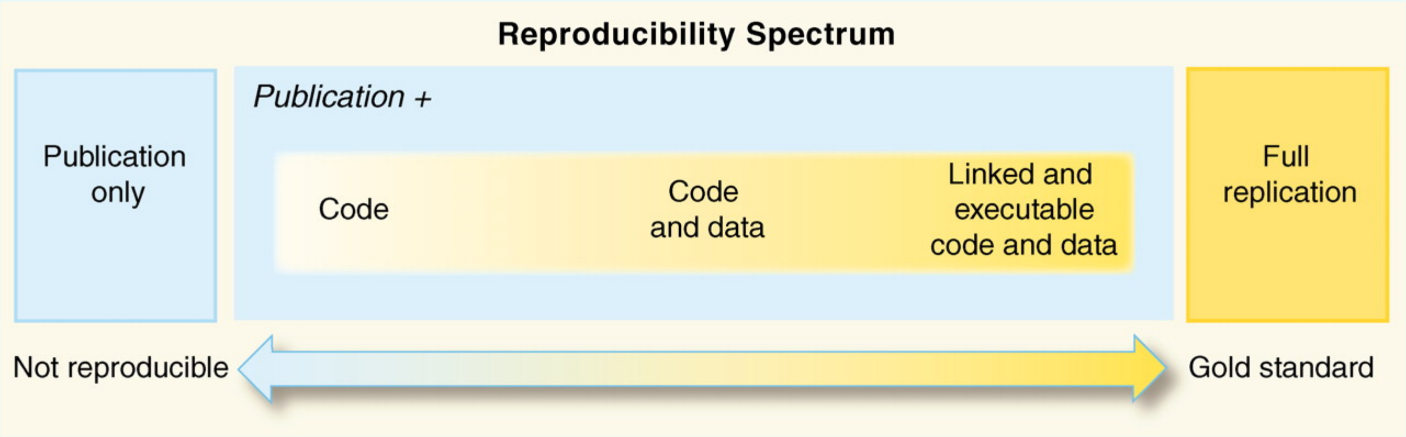
> “an article about a computational result is advertising, not scholarship. The actual scholarship is the full software environment, code and data, that produced the result.”

John Claerbout paraphrased in Buckheit and Donoho (1995)

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

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

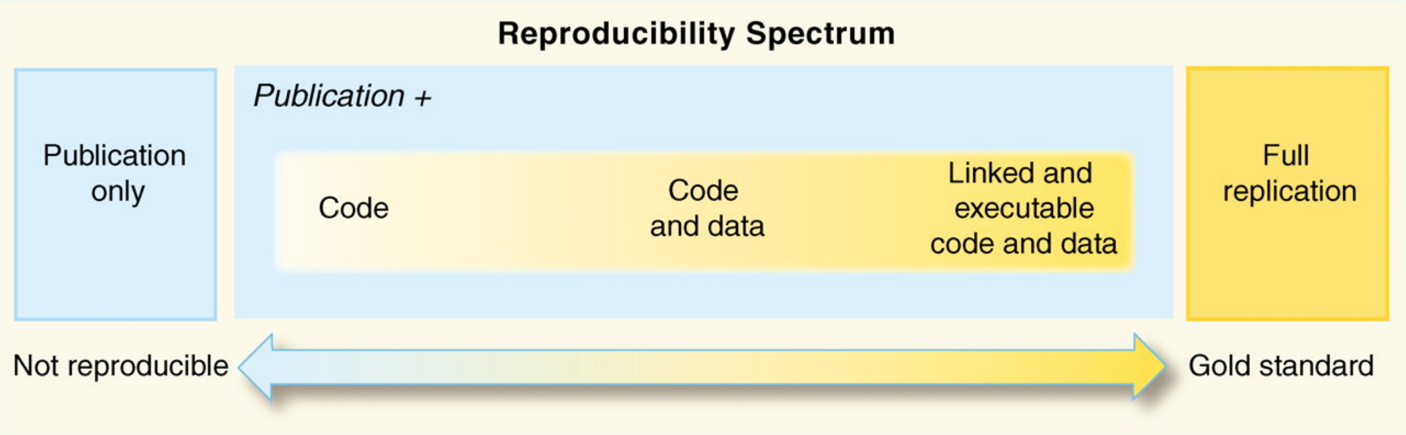
Benefit #1

transparency as a means of verification

Benefit #2

transparency as a means of supercharging research cycle

So how are we doing?



Reproducibility Spectrum

Publication only

Publication +

Code

Code and data

Linked and executable code and data

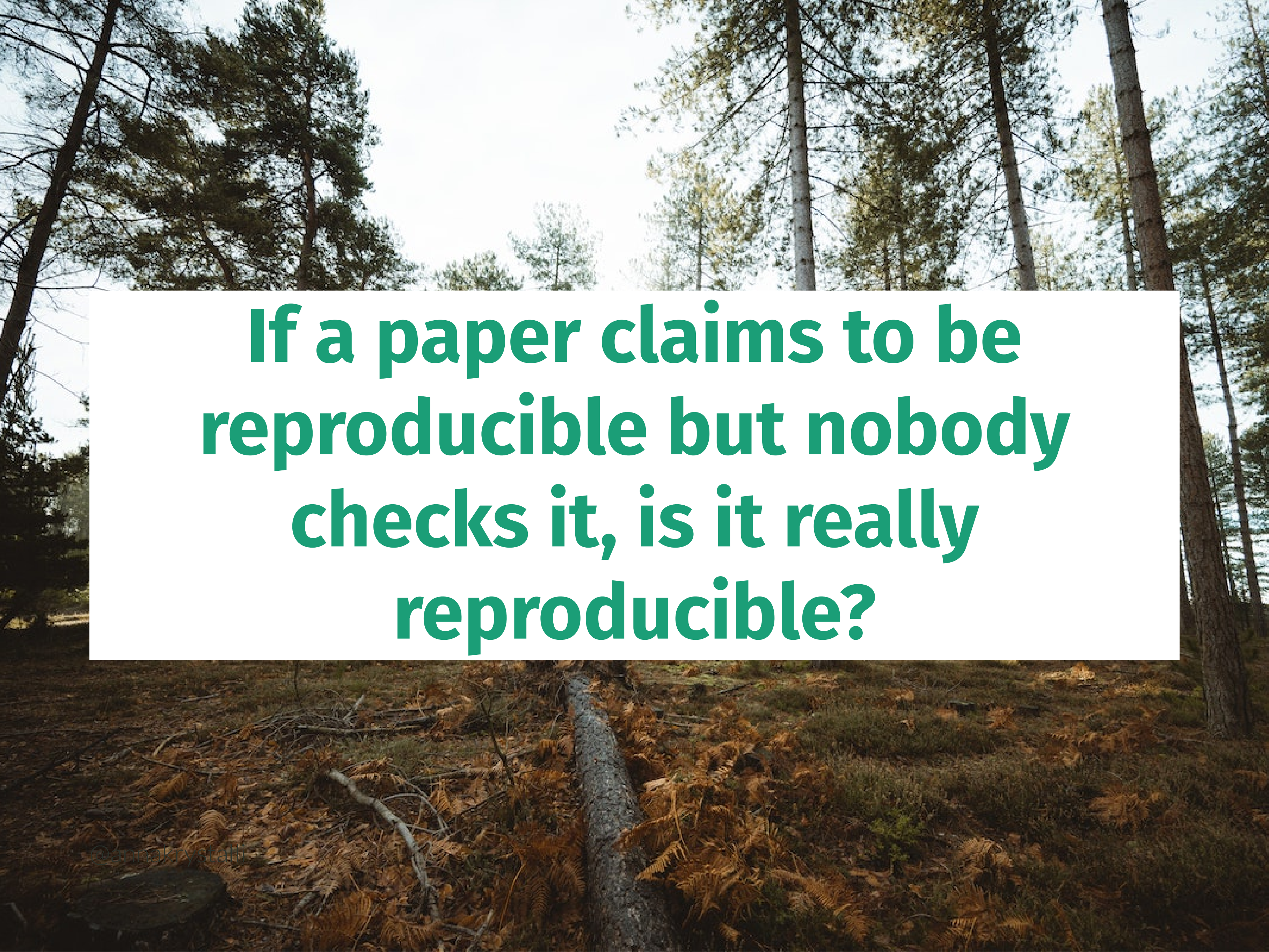
Full replication

Not reproducible

Gold standard

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

@annakrystalli



If a paper claims to be reproducible but nobody checks it, is it really reproducible?

@annakrystalli

Practice

Reprohack

One day reproducibility hackathons

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

ReproHack History

OpenCon Satellite: Berlin, 2016

OpenCon Satellite: London, 2017

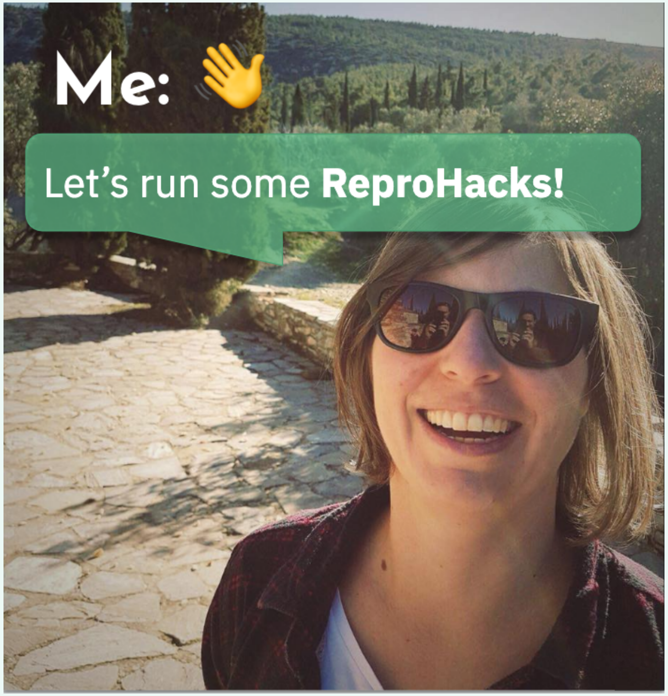
Inspired by Owen Petchey's Reproducible Research in Ecology, Evolution, Behaviour, and Environmental Studies course,

- 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



CCMcr19

CarpentryConnect Manchester 2019

@annakrystalli

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

Reprohack Core Team

How does it work?

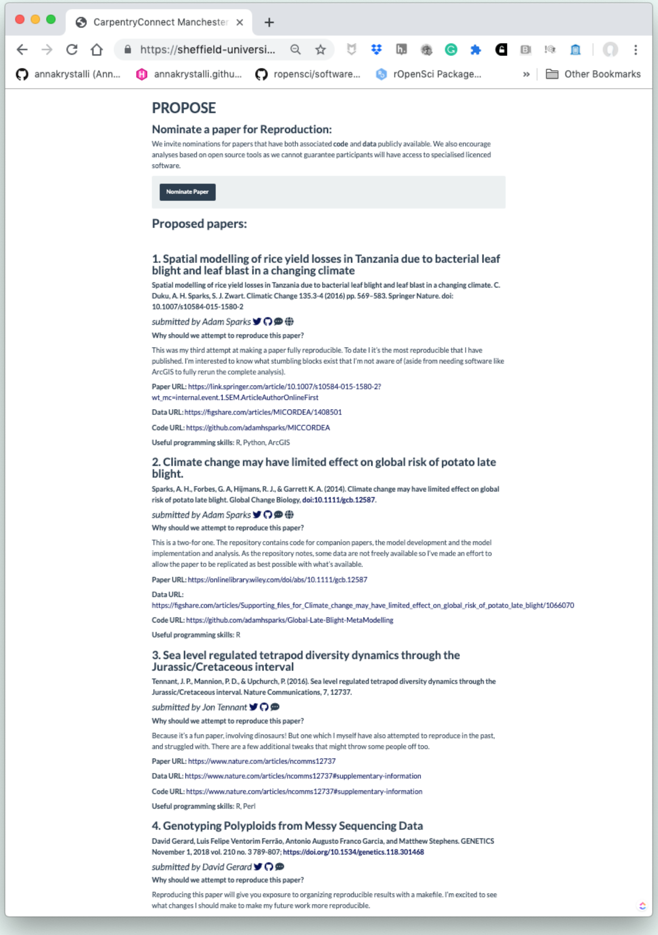
Call for papers

✨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/PssdXqwl8Z

— annakrystalli (@annakrystalli) June 12, 2019



PROPOSE

Nominate a paper for Reproduction:

We invite nominations for papers that have both associated **code** and **data** publicly available. We also encourage analyses based on open source tools as we cannot guarantee participants will have access to specialised licenced software.

Proposed papers:

1. Spatial modelling of rice yield losses in Tanzania due to bacterial leaf blight and leaf blast in a changing climate

Spatial modelling of rice yield losses in Tanzania due to bacterial leaf blight and leaf blast in a changing climate. C. Duku, A. H. Sparks, S. J. Zwart. Climatic Change 135.3-4 (2016) pp. 569–583. Springer Nature. doi: 10.1007/s10584-015-1580-2

submitted by Adam Sparks

Why should we attempt to reproduce this paper?

This was my third attempt at making a paper fully reproducible. To date I it's the most reproducible that I have published. I'm interested to know what stumbling blocks exist that I'm not aware of (aside from needing software like ArcGIS to fully rerun the complete analysis).

Paper URL: https://link.springer.com/article/10.1007/s10584-015-1580-2?wt_mc=internal.event.1.SEM.ArticleAuthorOnlineFirst

Data URL: https://figshare.com/articles/MICORDEA/1408501

Code URL: https://github.com/adamhsparks/MICCORDEA

Useful programming skills: R, Python, ArcGIS

2. Climate change may have limited effect on global risk of potato late blight.

Sparks, A. H., Forbes, G. A, Hijmans, R. J., & Garrett K. A. (2014). Climate change may have limited effect on global risk of potato late blight. Global Change Biology, doi:10.1111/gcb.12587.

submitted by Adam Sparks

Why should we attempt to reproduce this paper?

This is a two-for one. The repository contains code for companion papers, the model development and the model implementation and analysis. As the repository notes, some data are not freely available so I've made an effort to allow the paper to be replicated as best possible with what's available.

Paper URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/gcb.12587

Data URL:
https://figshare.com/articles/Supporting_files_for_Climate_change_may_have_limited_effect_on_global_risk_of_potato_late_blight/1066070

Code URL: https://github.com/adamhsparks/Global-Late-Blight-MetaModelling

Useful programming skills: R

3. Sea level regulated tetrapod diversity dynamics through the Jurassic/Cretaceous interval

Tennant, J. P., Mannion, P. D., & Upchurch, P. (2016). Sea level regulated tetrapod diversity dynamics through the Jurassic/Cretaceous interval. Nature Communications, 7, 12737.

submitted by Jon Tennant

Why should we attempt to reproduce this paper?

Because it's a fun paper, involving dinosaurs! But one which I myself have also attempted to reproduce in the past, and struggled with. There are a few additional tweaks that might throw some people off too.

Paper URL: https://www.nature.com/articles/ncomms12737

Data URL: https://www.nature.com/articles/ncomms12737#supplementary-information

Code URL: https://www.nature.com/articles/ncomms12737#supplementary-information

Useful programming skills: R, Perl

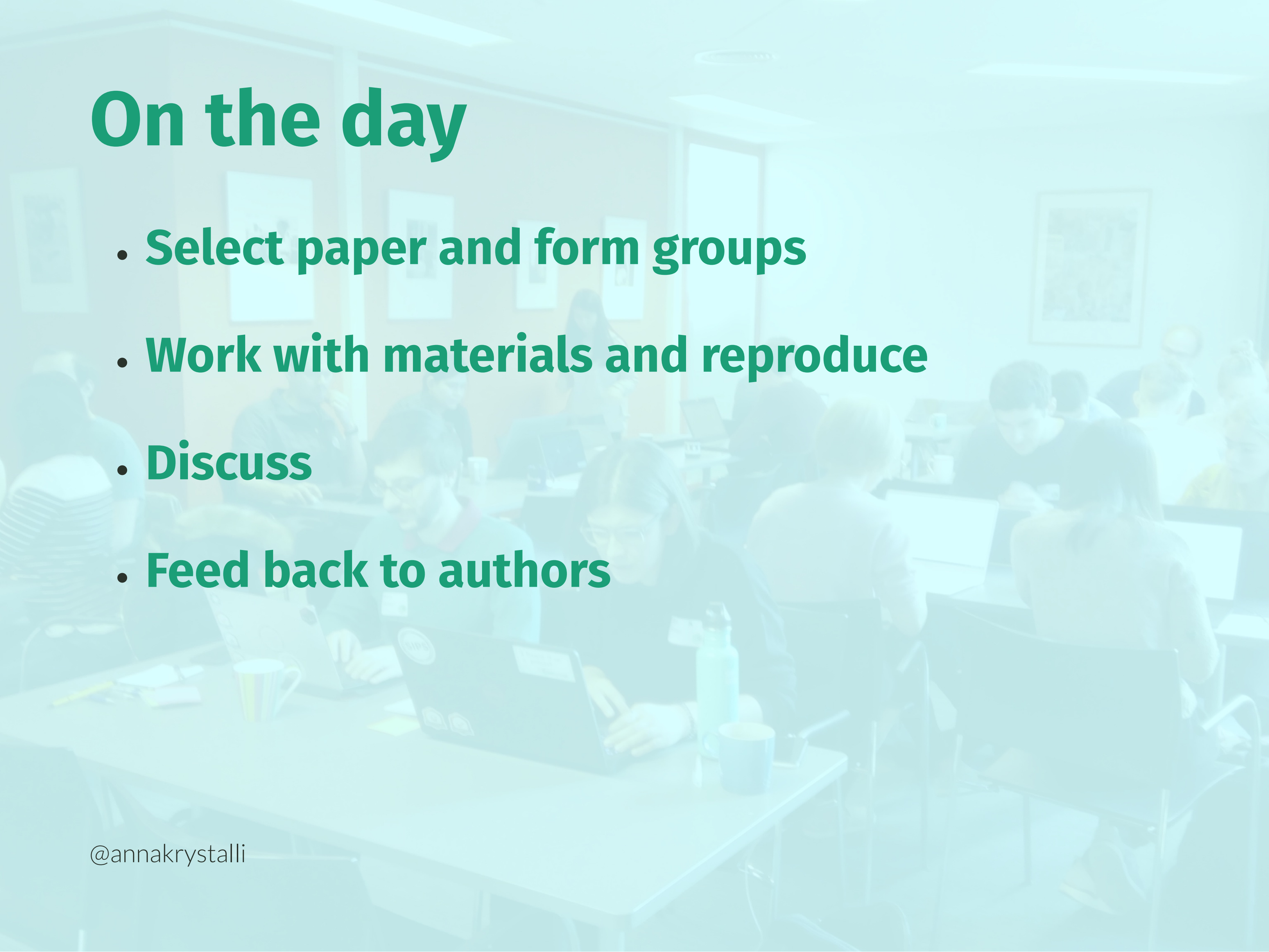
4. Genotyping Polyploids from Messy Sequencing Data

David Gerard, Luis Felipe Ventorim Ferrão, Antonio Augusto Franco Garcia, and Matthew Stephens. GENETICS November 1, 2018 vol. 210 no. 3 789-807; https://doi.org/10.1534/genetics.118.301468

submitted by David Gerard

Why should we attempt to reproduce this paper?

Reproducing this paper will give you exposure to organizing reproducible results with a makefile. I'm excited to see what changes I should make to make my future work more reproducible.



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 📑

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?

@annakrystalli

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?

@annakrystalli

Review as a user 🎮

New User

Invested User

Feedback as a community member

Acknowledge author effort

Give feedback in good faith

Focus on community benefits and system level solutions

@annakrystalli

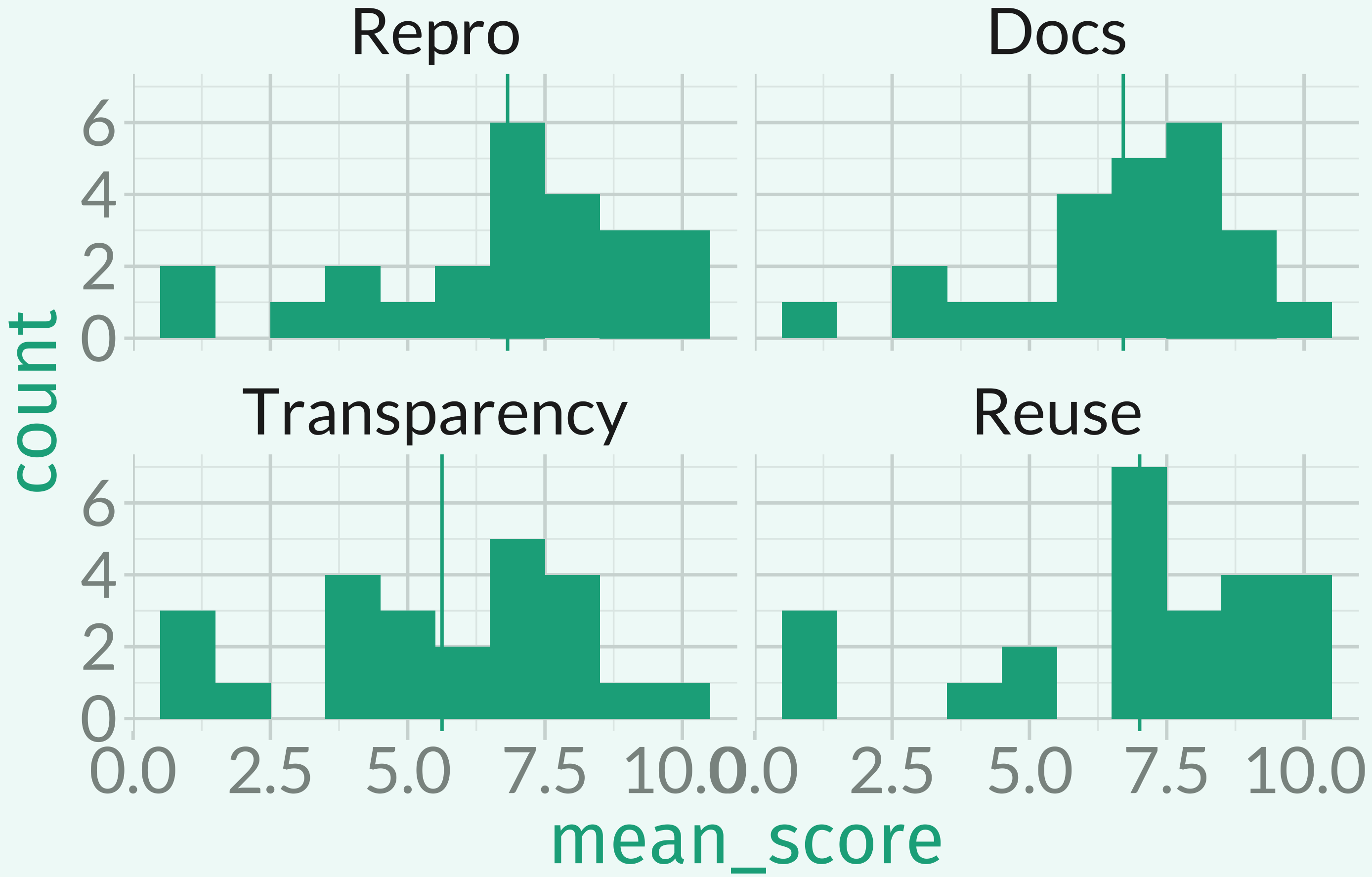
> *Help build convention on what form a Reproducible paper should take and how we should be able to use it*

What did we learn?

N8 CIR ReproHack Series Stats

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

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

ReproHacks are fun

On the future of Reviewing

REVIEWER

AUTHOR

I FOUND AND CORRECTED SOME TYPOS AND A MINOR [illegible] IN THE CODE. RESULTS [illegible]ED.

😃 THANKS!

I REANALYSED [illegible] PARAMETER X [illegible] SIGNIFICANT[illegible] RESULTS

WOW! THAT'S AN INTERESTING FINDING. IT WILL PROBABLY REQUIRE A SIGNIFICANT REWRITE. WOULD YOU LIKE TO COLLABORATE ON AN UPDATED MANUSCRIPT?

I THINK YOU SHOULD PERFORM COMPLICATED ANALYSIS Y

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

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/ CODE WORKS ✓

On the scope of reusability

Openness can help:

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

MAINTENANCE?!

…in the meantime

take any opportunity to practice!

ReproHack

Multiple ways to run a ReproHack

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.

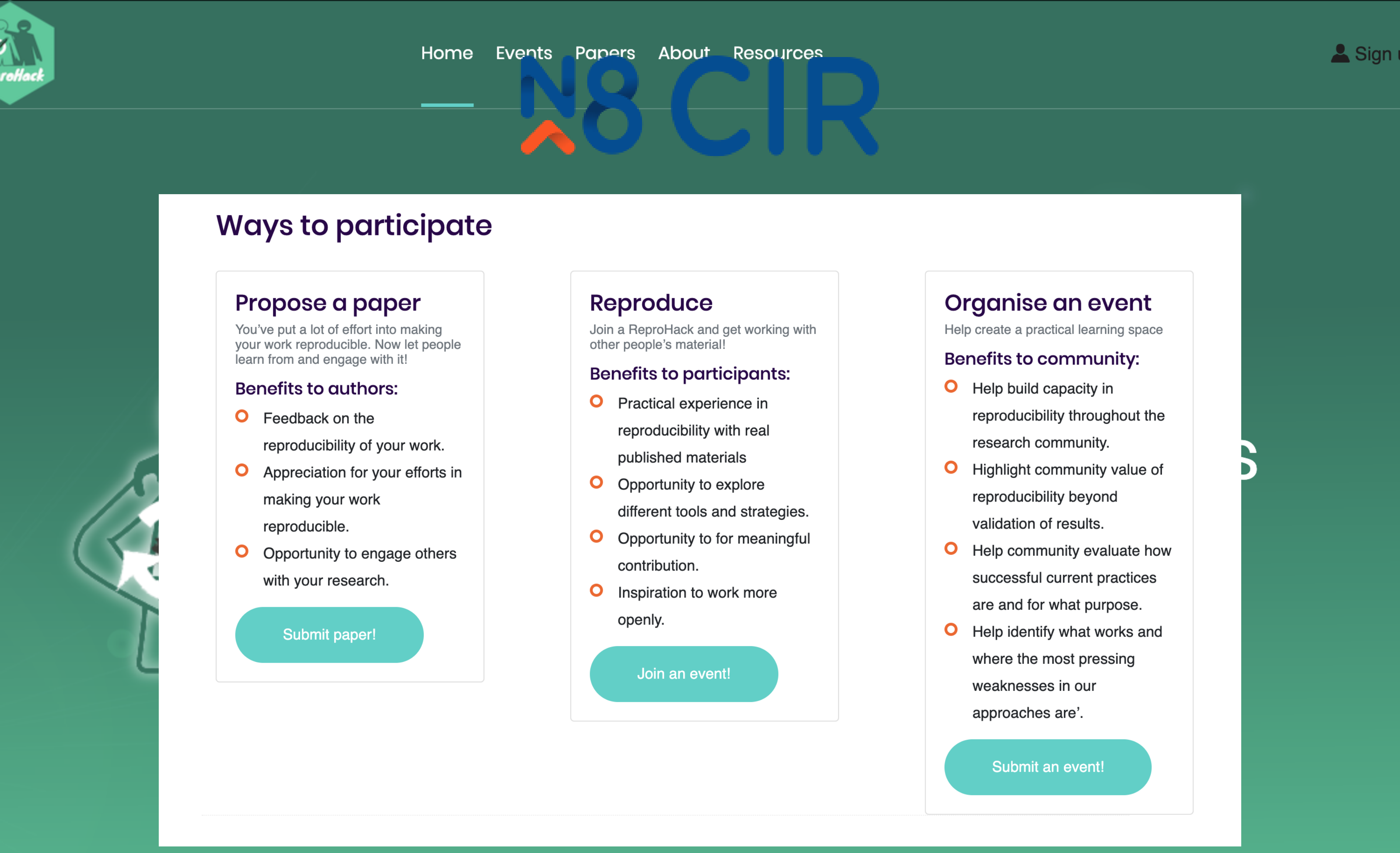
NO

Remote ReproHack

👩‍💻 👩‍💻 👨‍💻 | 👨‍💻 👩‍💻 | 👨‍💻

- ✓ 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.

@annakrystalli



N8 CIR

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!

@annakrystalli

Interested in ReproHacking?

reprohack/reprohack-hq GH repository

Chat to us:

slack join us

Host your own event!

Submit your own papers!

👋 Thanks for 👀

❓

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

- *The Turing Way Community, & Scriberia. (2019, July 11). Illustrations from the Turing Way book dashes. Zenodo. http://doi.org/10.5281/zenodo.3332808*

Photo on slide #24 by Annie Spratt on Unsplash

Photo on slide #25 Sharon McCutcheon on Unsplash