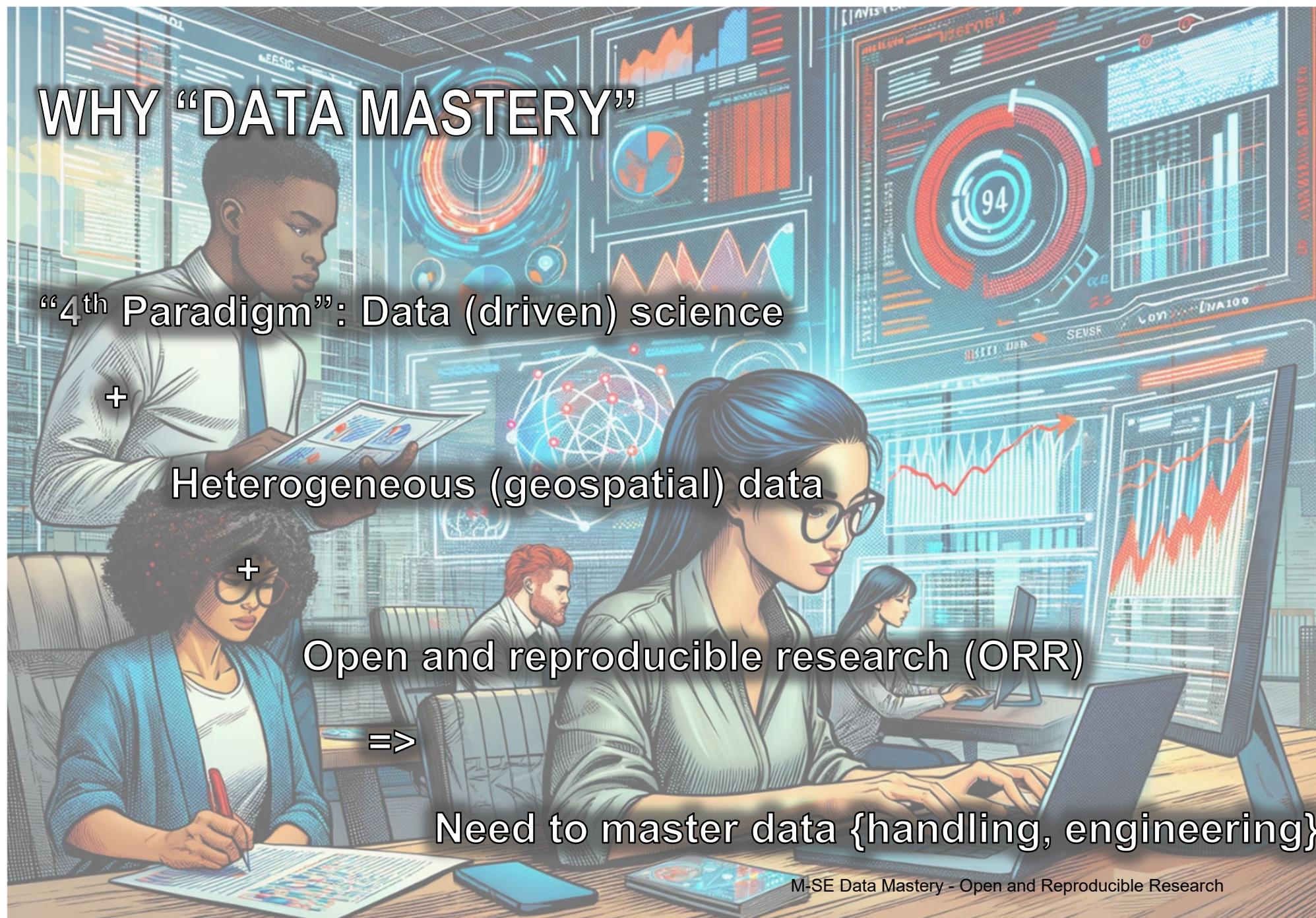


# Open, reproducible, and replicable research

Data Mastery Challenge

Frank Ostermann



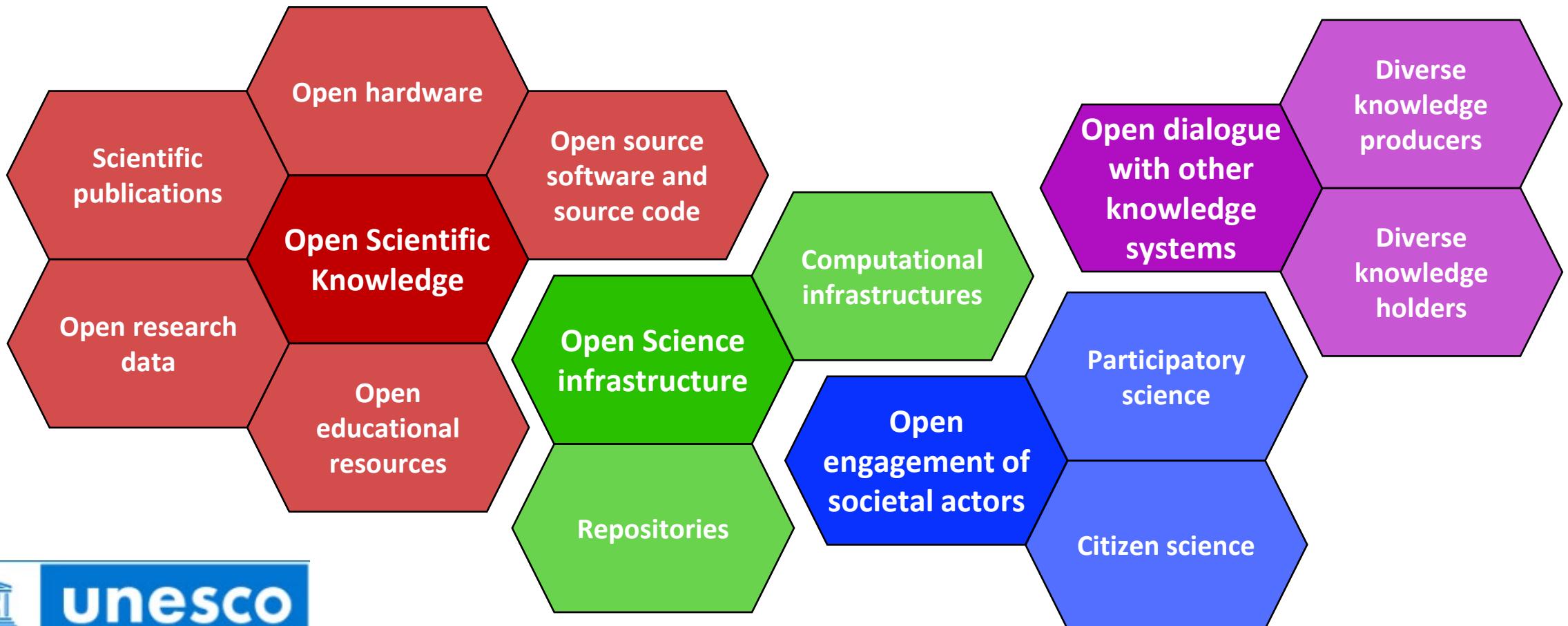


Source: DALL-E  
“create a comic  
style image  
showing data  
science in action”

# Part 1: Open Science and repr... repl... *What?*

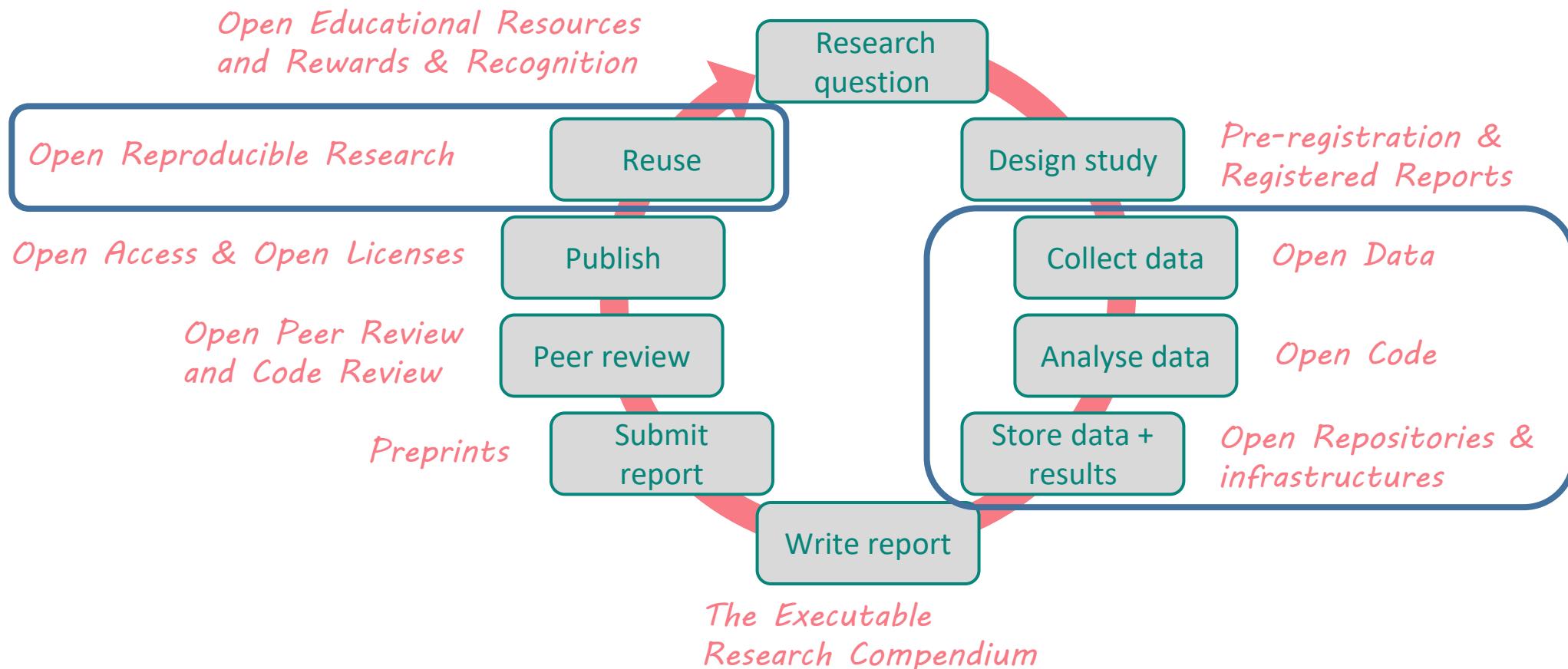
# What does Open Science mean?

---



unesco

# What does Open Science mean?





Sylvain Deville  
Freezing Observations in Earth Engineering  
Principles and Use  
Author of "Freezing Colloids"  
[springer.com/fr/book/9783319783311](http://springer.com/fr/book/9783319783311)

Sylvain   
@DevilleSy

Freezing stuff since 1876. Will science for chocolate. ORCID Id 0000-0002-3363-3184.

Author of "Freezing Colloids"  
[springer.com/fr/book/9783319783311](http://springer.com/fr/book/9783319783311)

France  
[sylvaindeville.net](http://sylvaindeville.net)

Sylvain   
@DevilleSy

Follow

When you try to replicate a paper using the methods section

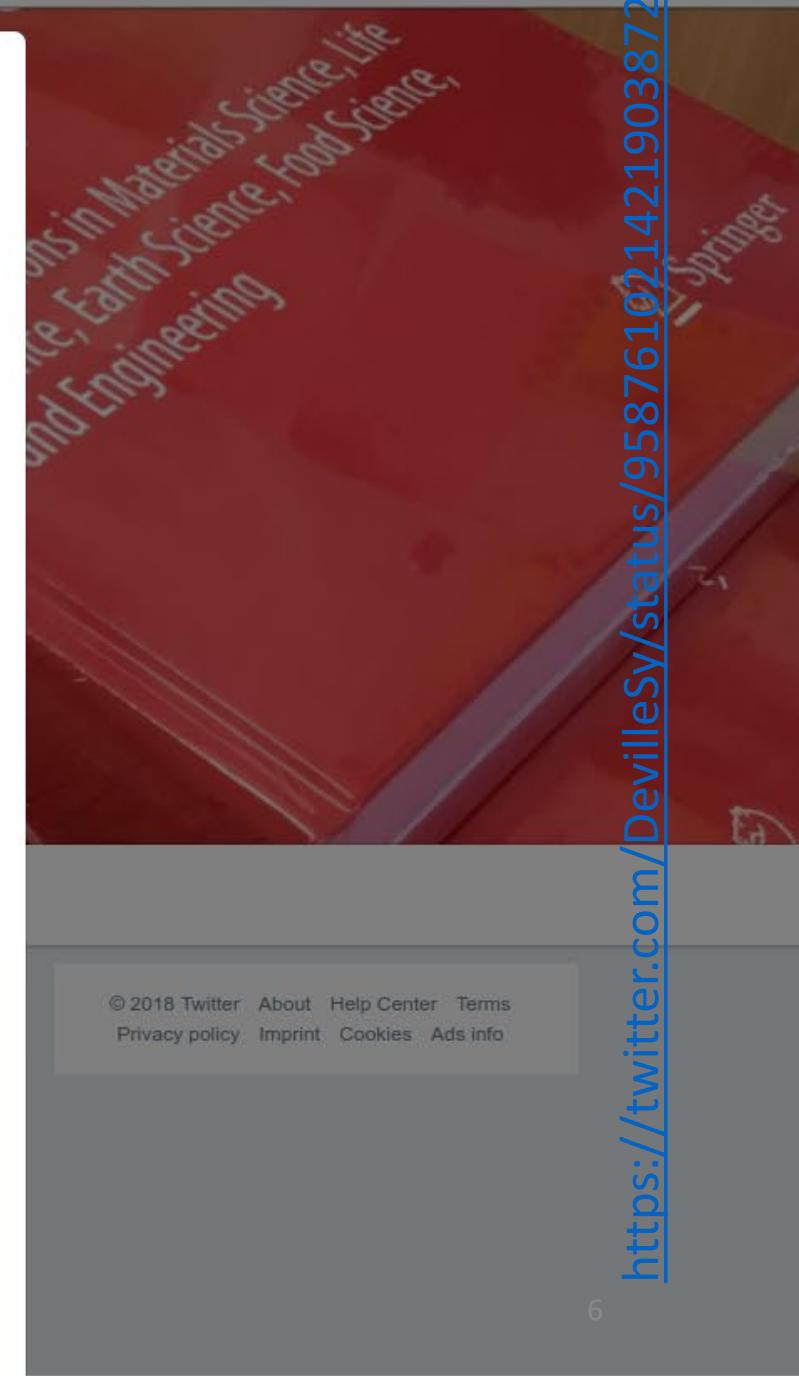


9:56 AM - 31 Jan 2018

2,605 Retweets 5,877 Likes

M-SE Data Mastery - Open and Reproducible Research

54 2.6K 5.9K



© 2018 Twitter About Help Center Terms  
Privacy policy Imprint Cookies Ads info

<https://twitter.com/DevilleSy/status/958761021421903872>

# Much ado about reproducibility?

PLoS Med. 2005 Aug; 2(8): e124.

Published online 2005 Aug 30. doi: [10.1371/journal.pmed.0020124](https://doi.org/10.1371/journal.pmed.0020124)

## Why Most Published Research Findings Are False

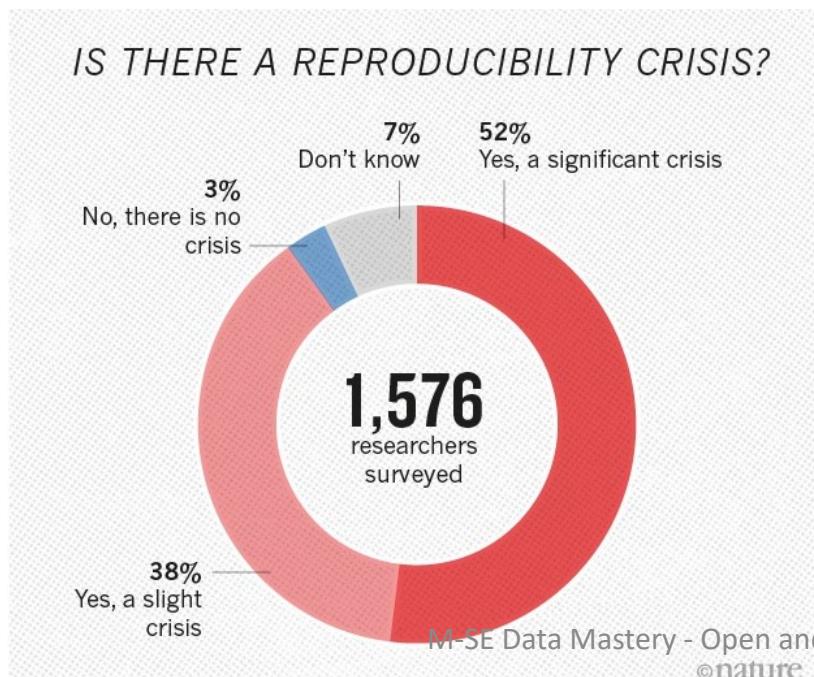
[John P. A. Ioannidis](#)

Published: 25 May 2016

## 1,500 scientists lift the lid on reproducibility

Monya Baker

Nature 533, 452–454 (2016) |



## Science has been in a “replication crisis” for a decade. Have we learned anything?

Bad papers are still published. But some other things might be getting better.

By Kelsey Piper | Oct 14, 2020, 12:20pm EDT

<https://www.vox.com/future-perfect/21504366/science-replication-crisis-peer-review-statistics>

Matters Arising | Published: 14 October 2020

## Transparency and reproducibility in artificial intelligence

Benjamin Haibe-Kains George Alexandru Adam, Ahmed Hosny, Farnoosh Khodakarami, Massive Analysis Quality Control (MAQC) Society Board of Directors, Levi Waldron, Bo Wang, Chris McIntosh, Anna Goldenberg, Anshul Kundaje, Casey S. Greene, Tamara Broderick, Michael M. Hoffman, Jeffrey T. Leek, Keegan Korthauer, Wolfgang Huber, Alvis Brazma, Joelle Pineau, Robert Tibshirani, Trevor Hastie, John P. A. Ioannidis, John Quackenbush & Hugo J. W. L. Aerts

Nature 586, E14–E16 (2020) | Cite this article

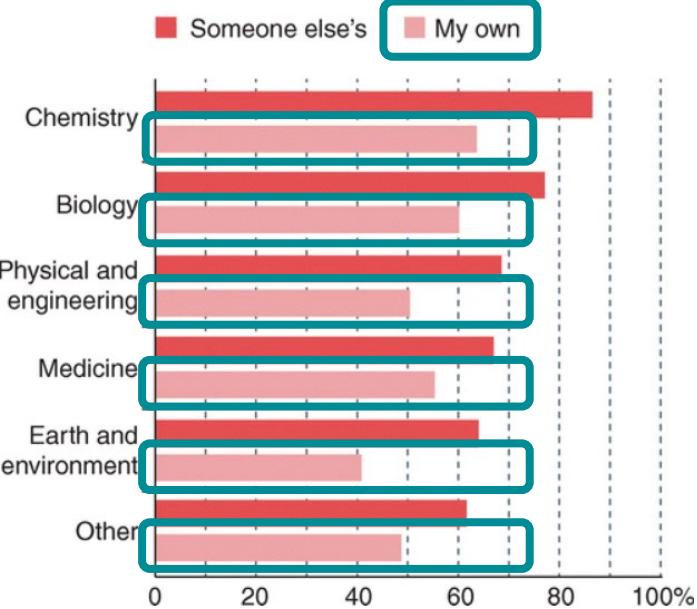
by Will Douglas Heaven  
November 12, 2020



Tech giants dominate research but the line between real breakthrough and product showcase can be fuzzy. Some scientists have had enough.

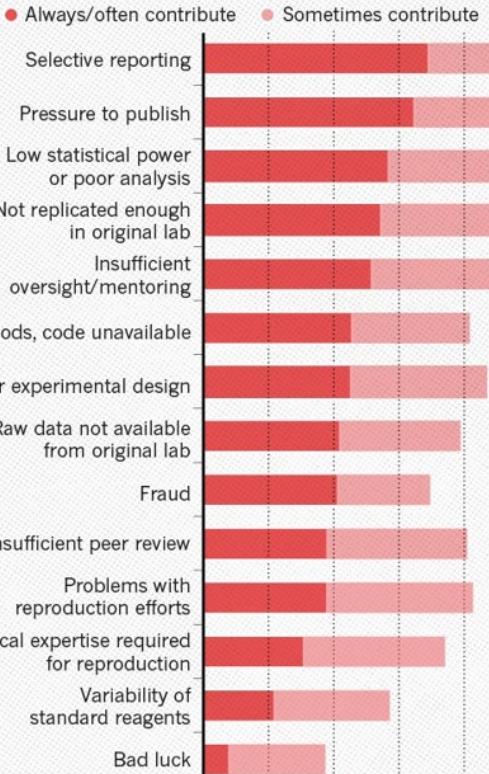
# The Reproducibility Crisis

Most scientists have experienced failure to reproduce results



## WHAT FACTORS CONTRIBUTE TO IRREPRODUCIBLE RESEARCH?

Many top-rated factors relate to intense competition and time pressure.



## WHAT FACTORS COULD BOOST REPRODUCIBILITY?

Respondents were positive about most proposed improvements but emphasized training in particular.



# Introduction to Open Reproducible Research

		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

## REPRODUCIBILITY DEFINITION

### REPRODUCIBILITY

“**Reproducibility** is the ability to get the same research results using the raw data and computer programs provided by the researchers. A related concept is **replicability**, meaning the ability to independently achieve similar conclusions when differences in sampling, research procedures and data analysis methods may exist.” (Wikipedia)

# What are reproducibility and replicability?

**Reproducibility** means the same results or outcomes when

- using the same original data
- applying the same methods (code, libraries, programs)

If outcomes are identical or within the expected margin of error: great, the original hypothesis has not been falsified, and research design is sound

**Replicability** means changing

- input data (time, geographic area, means of collections, etc.)
- methods (different libraries or completely different algorithm)

If outcomes are similar, original hypothesis is supported

If not, original hypothesis is not automatically falsified, but at least of limited generalizability (and if multiple replications fail, probably just an idiographic observation)

# What was my original motivation?

## Working with geosocial media / VGI:

- Platform (API) Black boxes: You can't guarantee that others will retrieve the same data
- Volatility of content and access: You can't guarantee that the content will remain the same, nor that others will continue to be able to access it (licenses, ToS)
- Variance in human behavior (inter- and intra-rater agreement): You can't guarantee that volunteer data is consistent, even from one participant

[Explore this journal >](#)

### Research Article

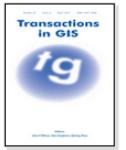
#### Advancing Science with VGI: Reproducibility and Replicability of Recent Studies using VGI

Frank O. Ostermann  [Carlos Granell](#)

First published: 5 January 2016 [Full publication history](#)

DOI: [10.1111/tgis.12195](https://doi.org/10.1111/tgis.12195) [View/save citation](#)

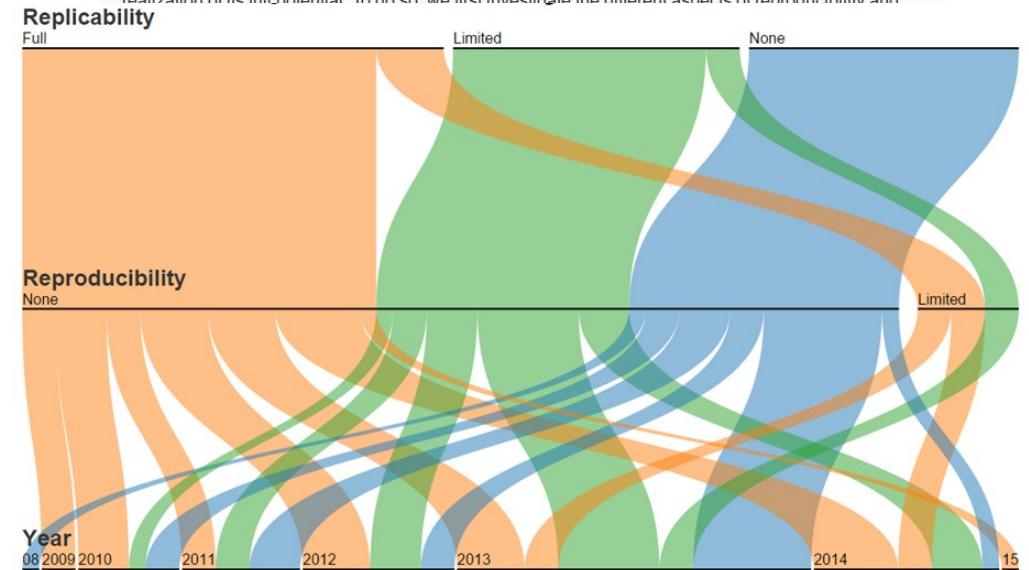
Cited by (CrossRef): 0 articles  [Citation tools](#)  2



[View issue TOC](#)  
Volume 21, Issue 2  
April 2017  
Pages 224–237

### Abstract

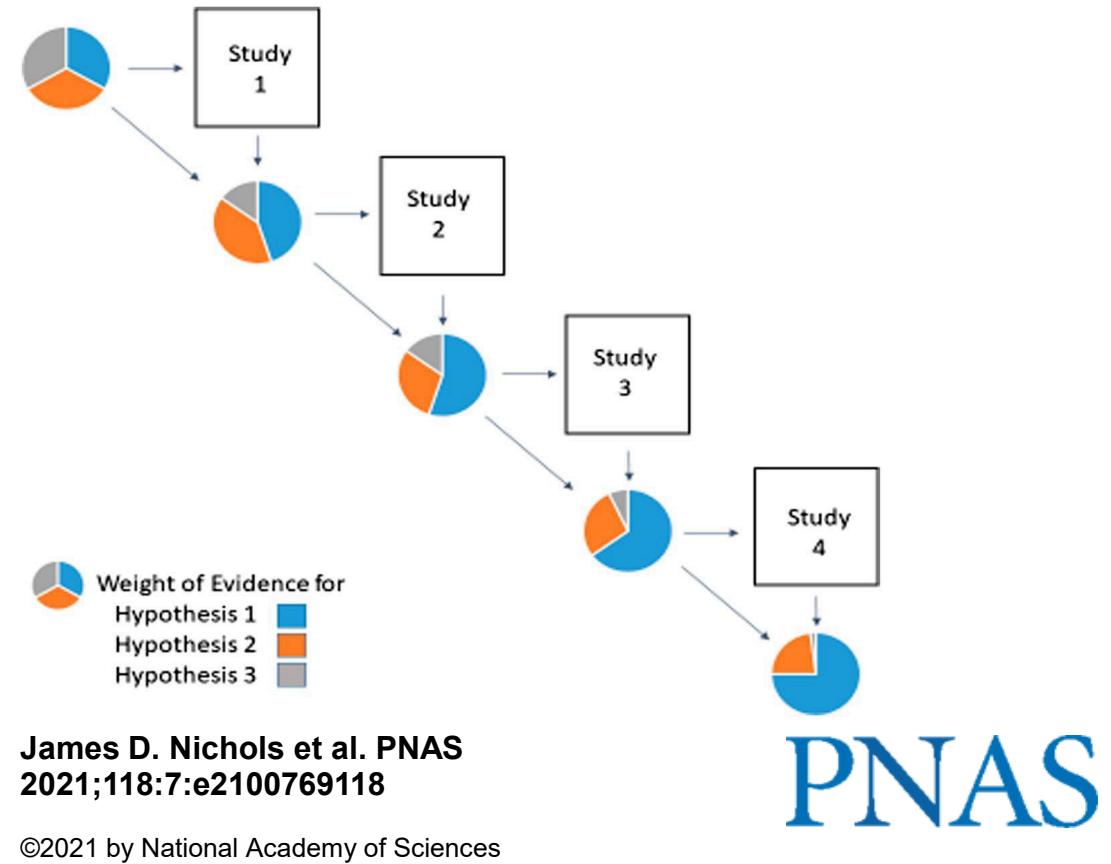
In scientific research, reproducibility and replicability are requirements to ensure the advancement of our body of knowledge. This holds true also for VGI-related research and studies. However, the characteristics of VGI suggest particular difficulties in ensuring reproducibility and replicability. In this article, we aim to examine the current situation in VGI-related research, and identify strategies to ensure realization of its full potential. To do so, we first investigate the different aspects of reproducibility and

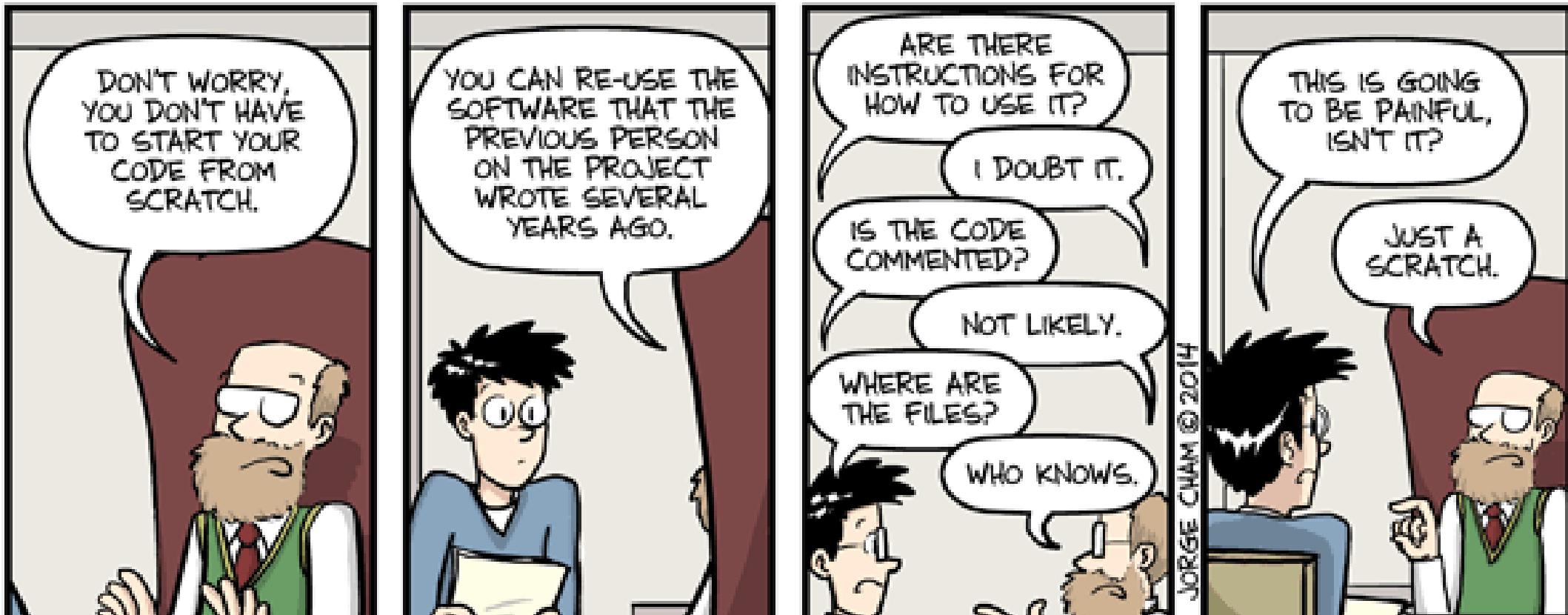


# Why do they matter?

For (open) science: Discover laws, axioms, rules, etc. and describe them and under which condition they apply

- Without reproducibility, replication is difficult (if you don't know which factors you changed, how can you interpret the new results?)
- Without replication, limited new knowledge (how do you know which observations are generalizable under which conditions?)





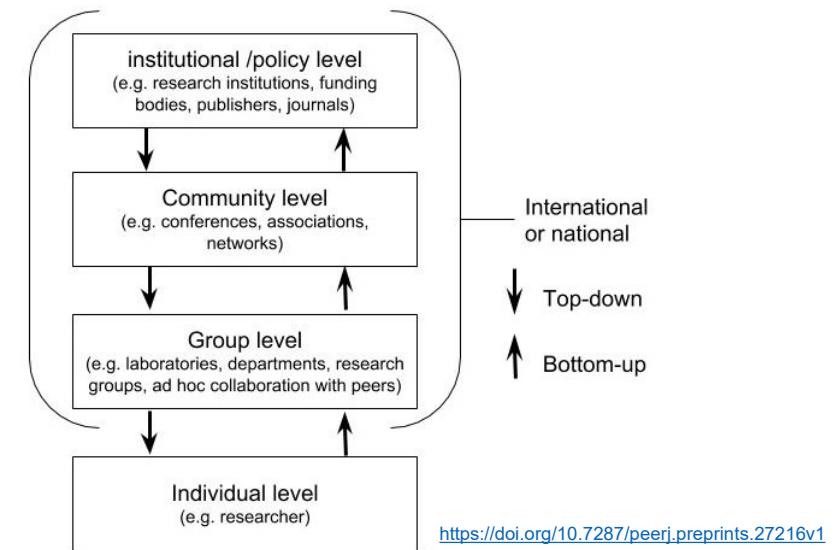
JORGE CHAM © 2014

[WWW.PHDCOMICS.COM](http://phdcomics.com/comics.php?f=1689)

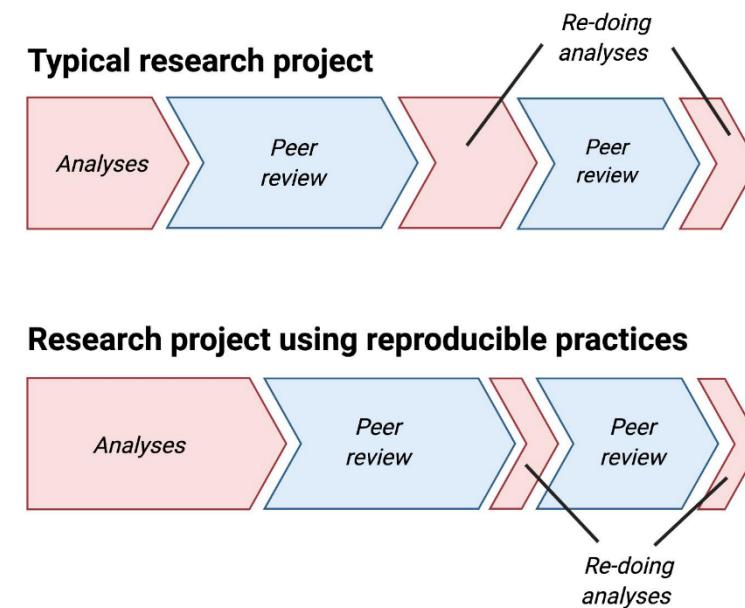
# Why do they matter?

For individual actors (YOU!):

- helps to avoid disaster of re-doing entire analyses
- makes it easier to write papers
- helps reviewers see it your way
- enables continuity of your work
- helps to build your reputation

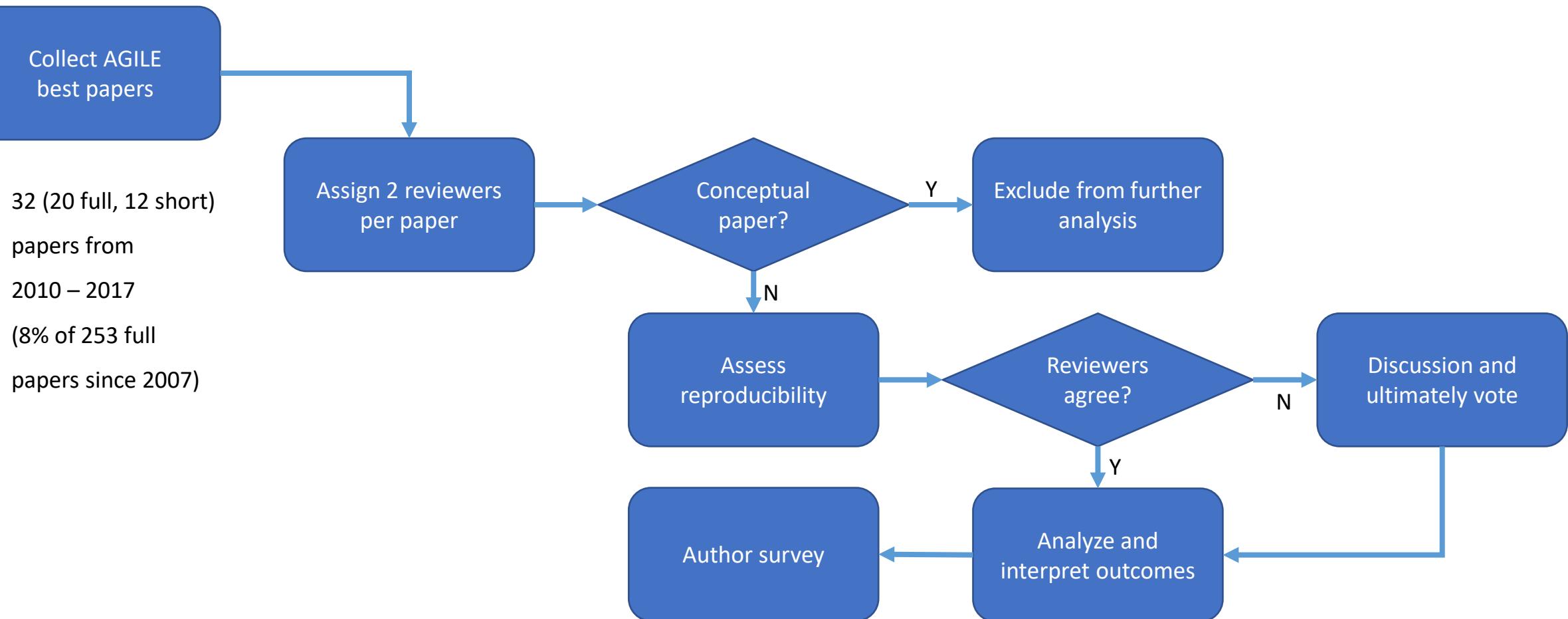


<https://doi.org/10.7287/peerj.preprints.27216v1>



Quintana, D. S. (2020, November 28). Five things about open and reproducible science that every early career researcher should know. <https://doi.org/10.17605/OSF.IO/DZTVQ>

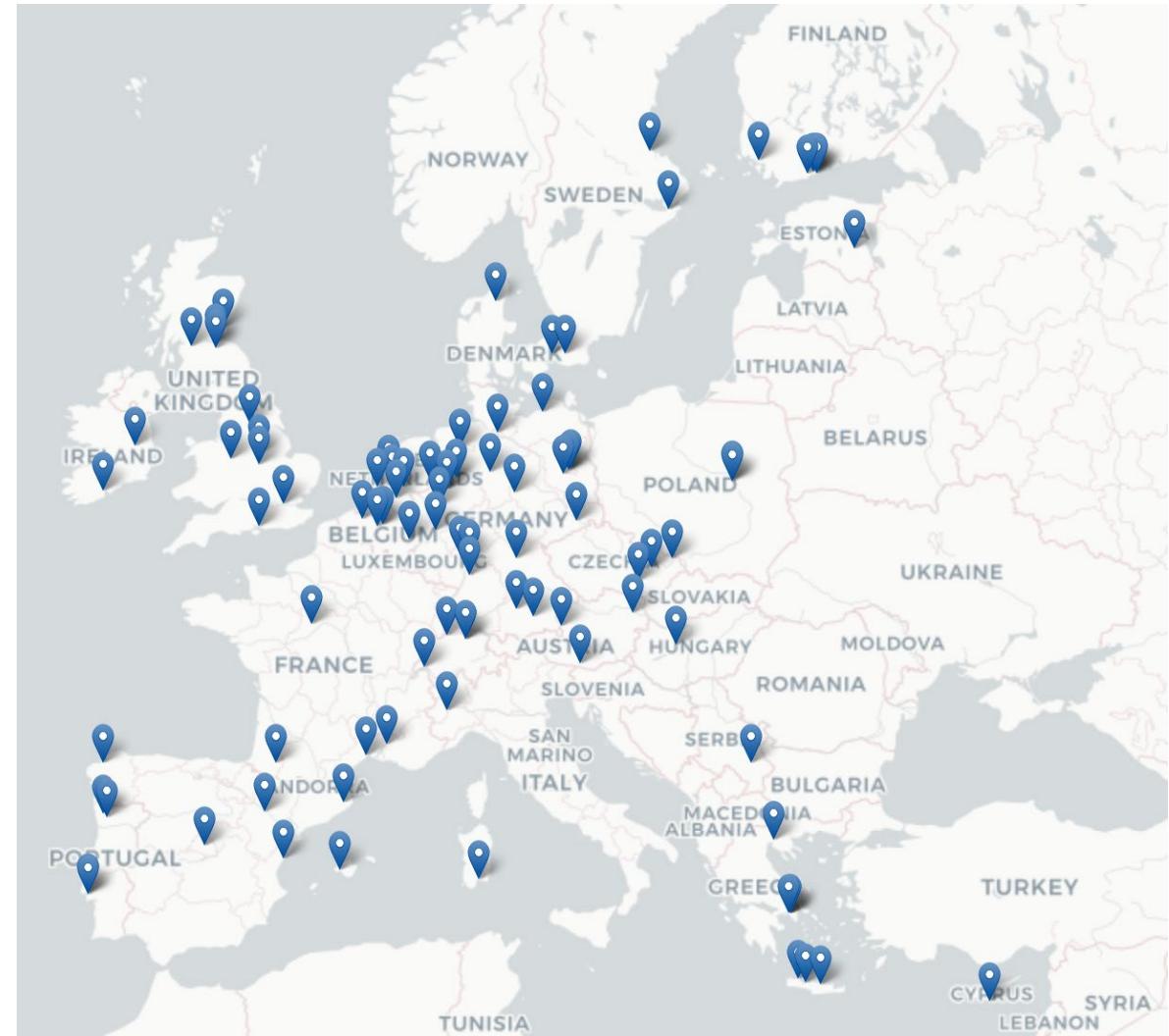
# Review paper 2017/18: How did we examine AGILE papers' reproducibility?



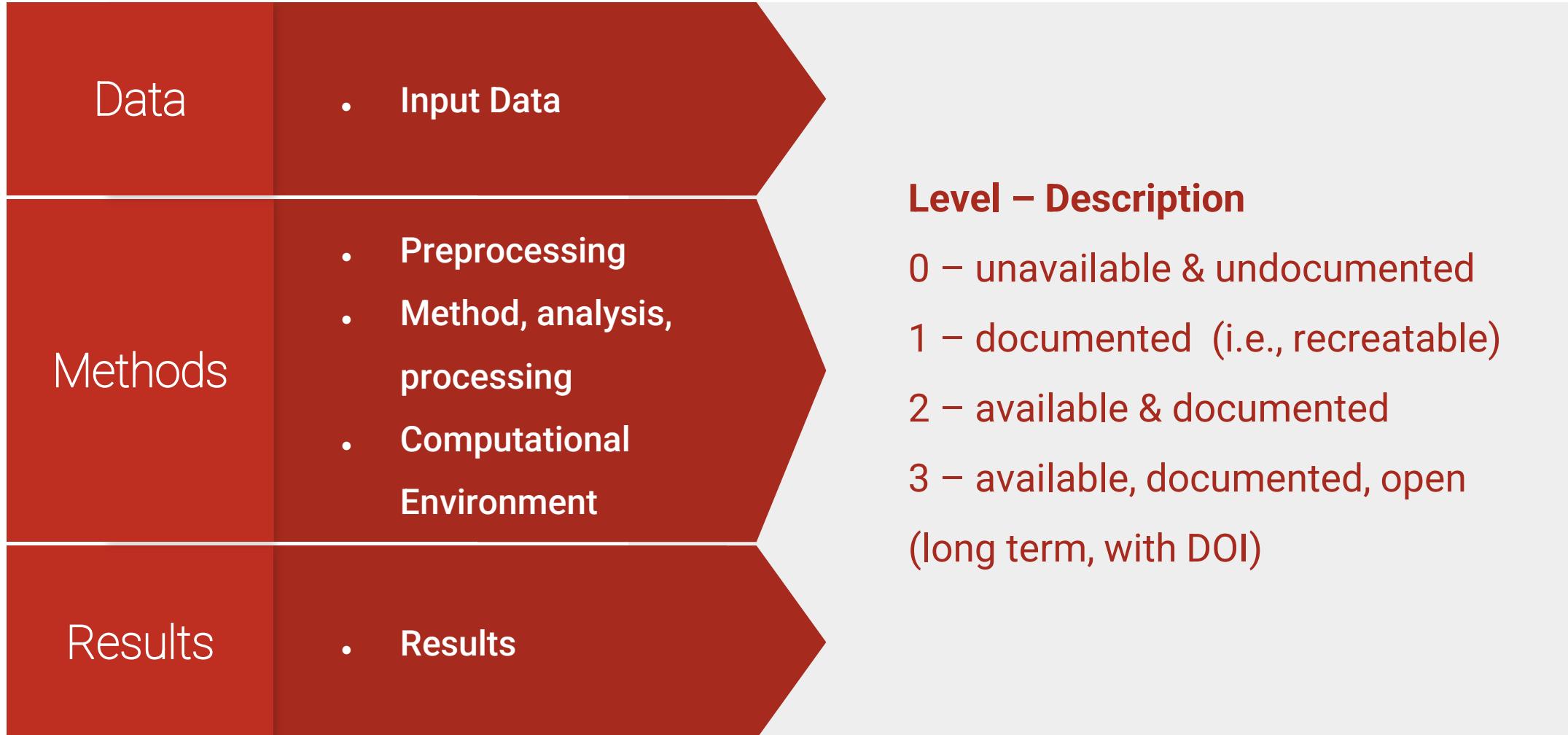
# Wait, what's AGILE?

**Association of Geographic Information Laboratories in Europe (<https://agile-online.org/>)**

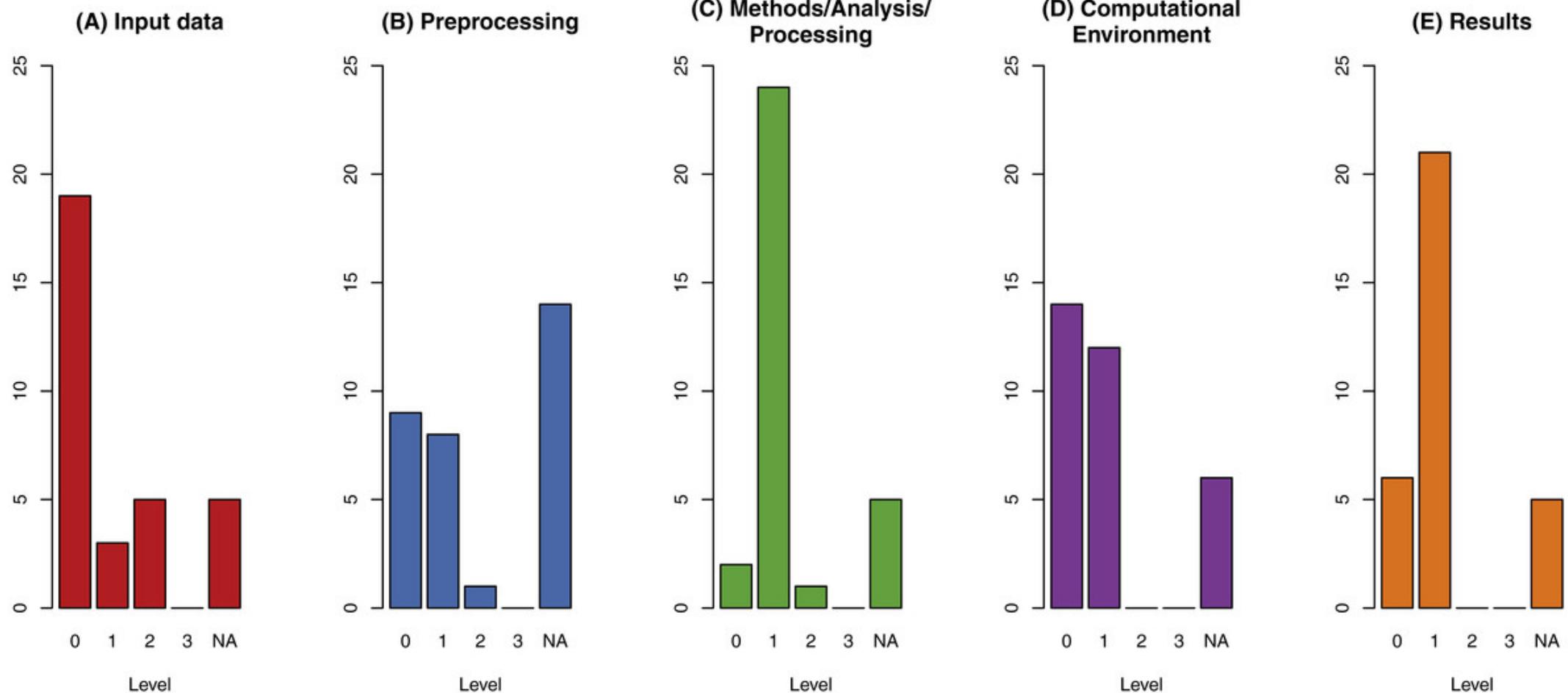
- Annual AGILE conference (<https://agile-online.org/conference-2020>)
- Bi-annual PhD School (<https://agile-online.org/agile-actions/phd-school>)
- AGILE Initiatives (<https://agile-online.org/funding-initiatives>)
- Collaboration & MoU with organizations & sister associations (<https://agile-online.org/agile-community/cooperation>)



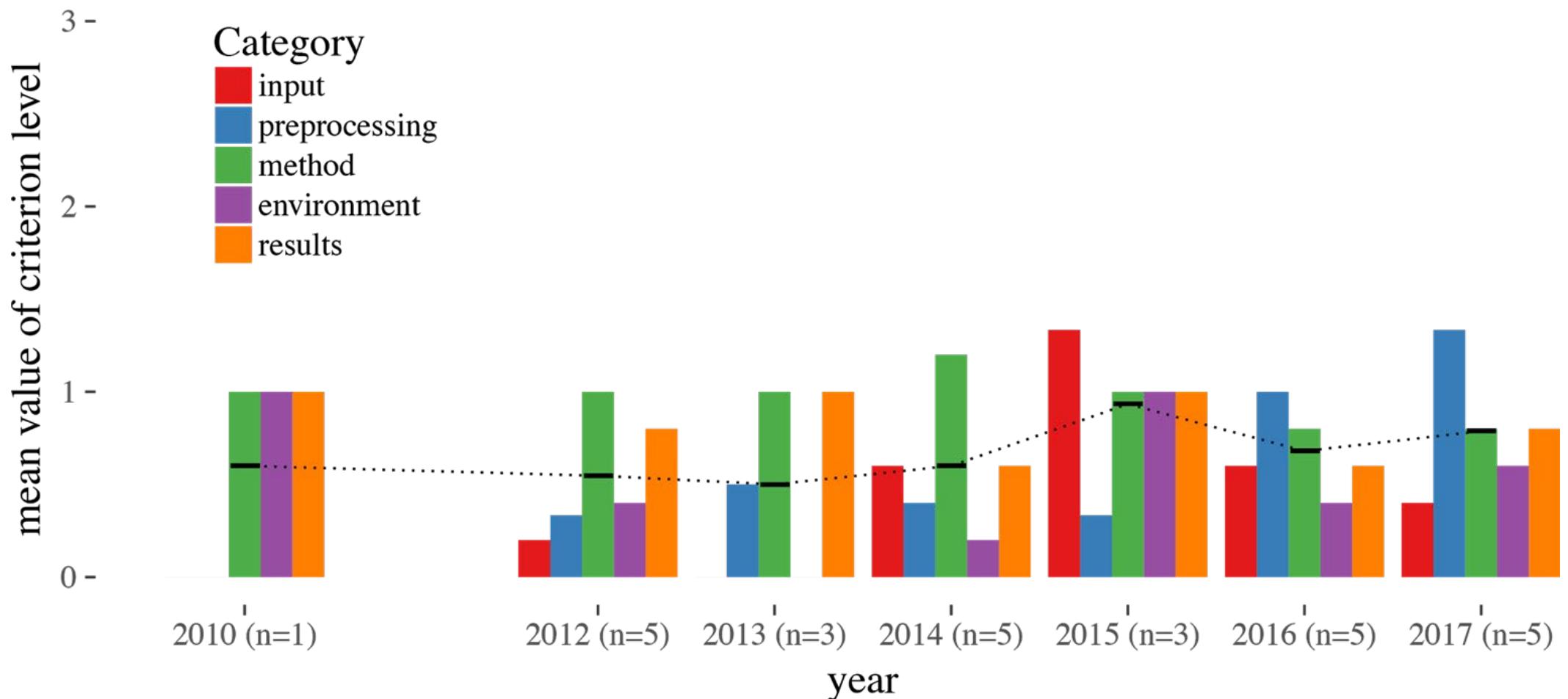
# How to assess reproducibility?



# How reproducible were AGILE papers?



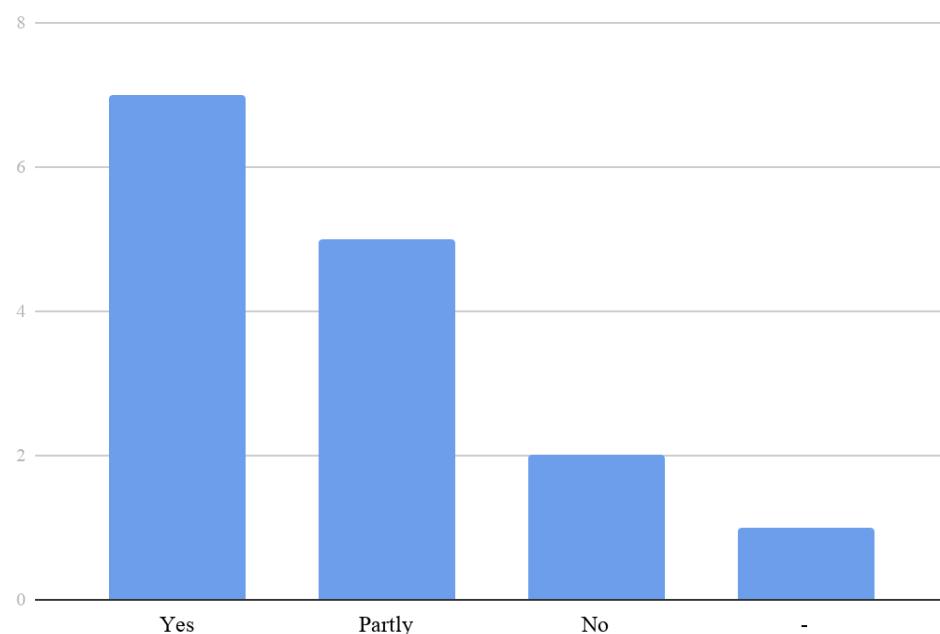
# Does it at least improve over time? (no)



# What were the authors' views?

- authors were provided with our evaluation of their paper
- 22 / 82 authors filled in the survey for 17/32 papers
- authors were asked to give consent to use their answers in the publications

**Do you agree with our assessment?**

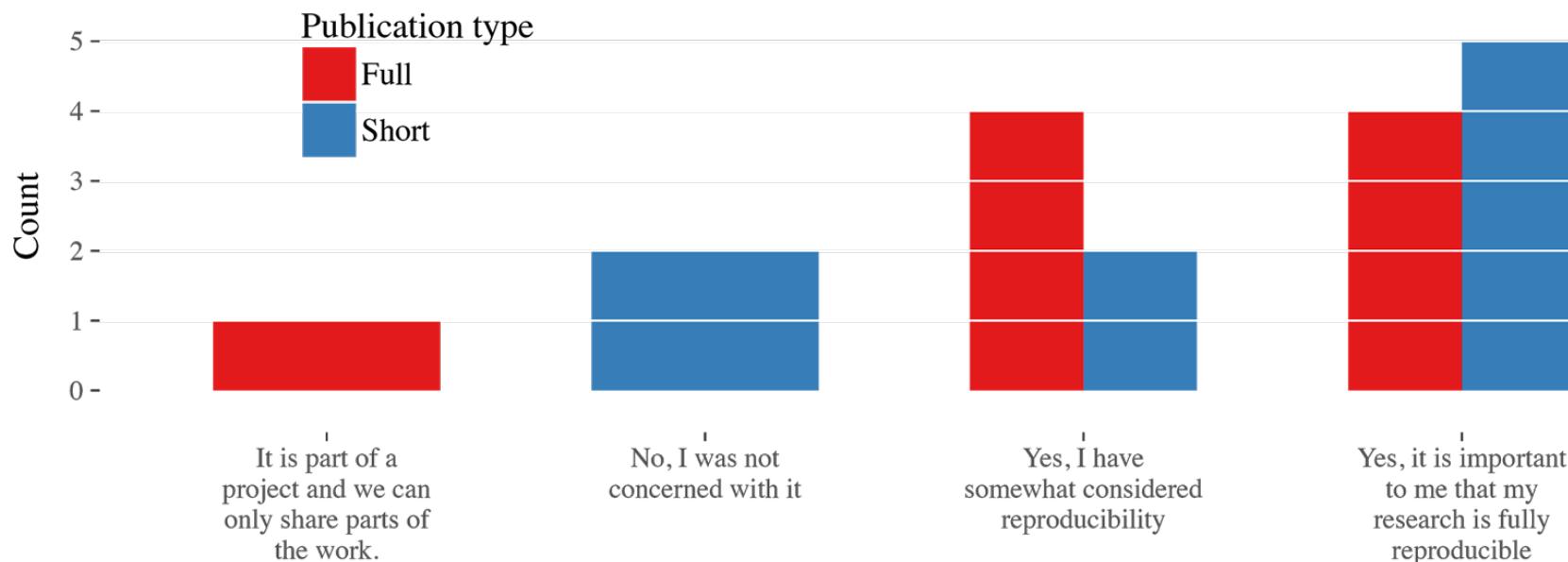


Reasons for disagreement:

- Requirements should not be applicable for short paper
- Specific data is not always necessary for reproducibility
- “Availability upon request” means “available”
- OSM data is by default “open and permanent”

# Did they consider reproducibility? Why not?

***Have you considered the reproducibility of research published in your nominated paper?***



## ***Reasons for lack of reproducibility***

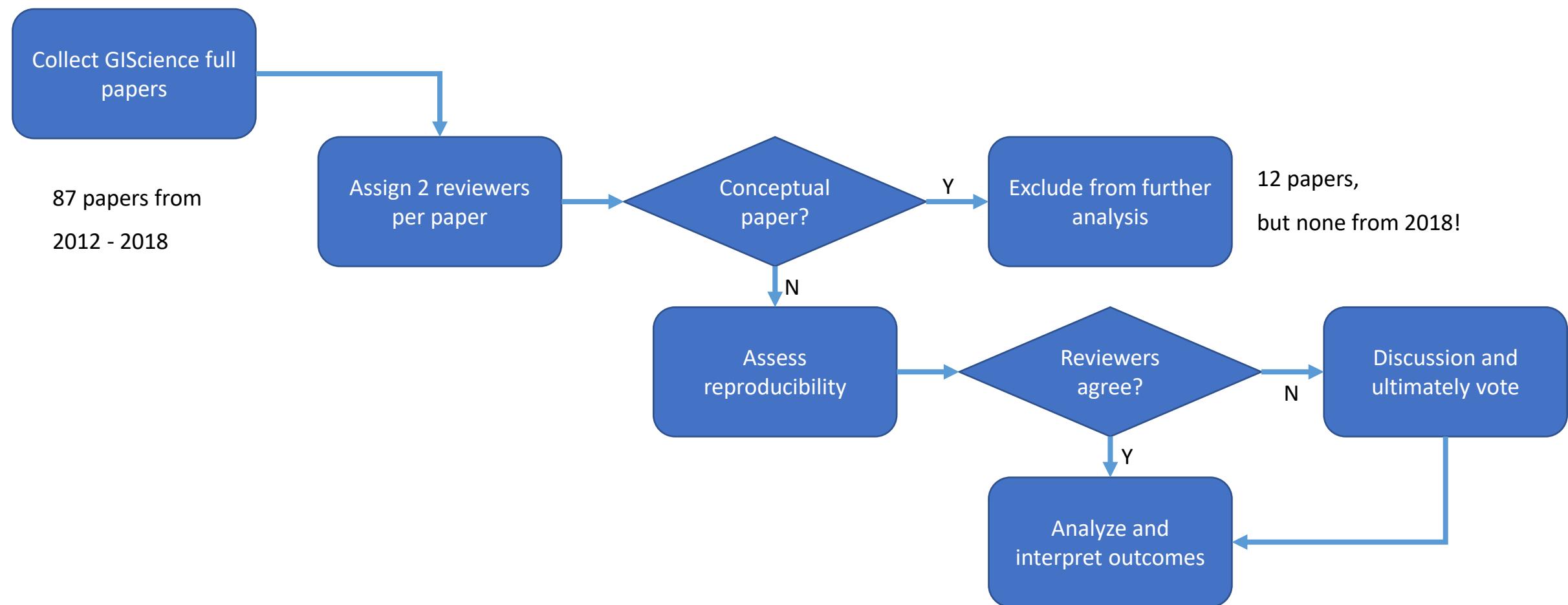
- Legal restrictions
- Not enough time
- Inadequate tools
- Lack of knowledge or skills
- Insufficient incentives

# Next step: Replicate with another conference

What did we want to do?

1. Investigate the state of reproducibility at GIScience conference series
2. Replicate an earlier assessment for AGILE conference series:
  - Is the method generalizable?
  - How do AGILE and GIScience compare?
3. Discuss strategies for improving reproducibility

# How did we go about it?



# Was our approach replicable?

Short answer: yes

But:

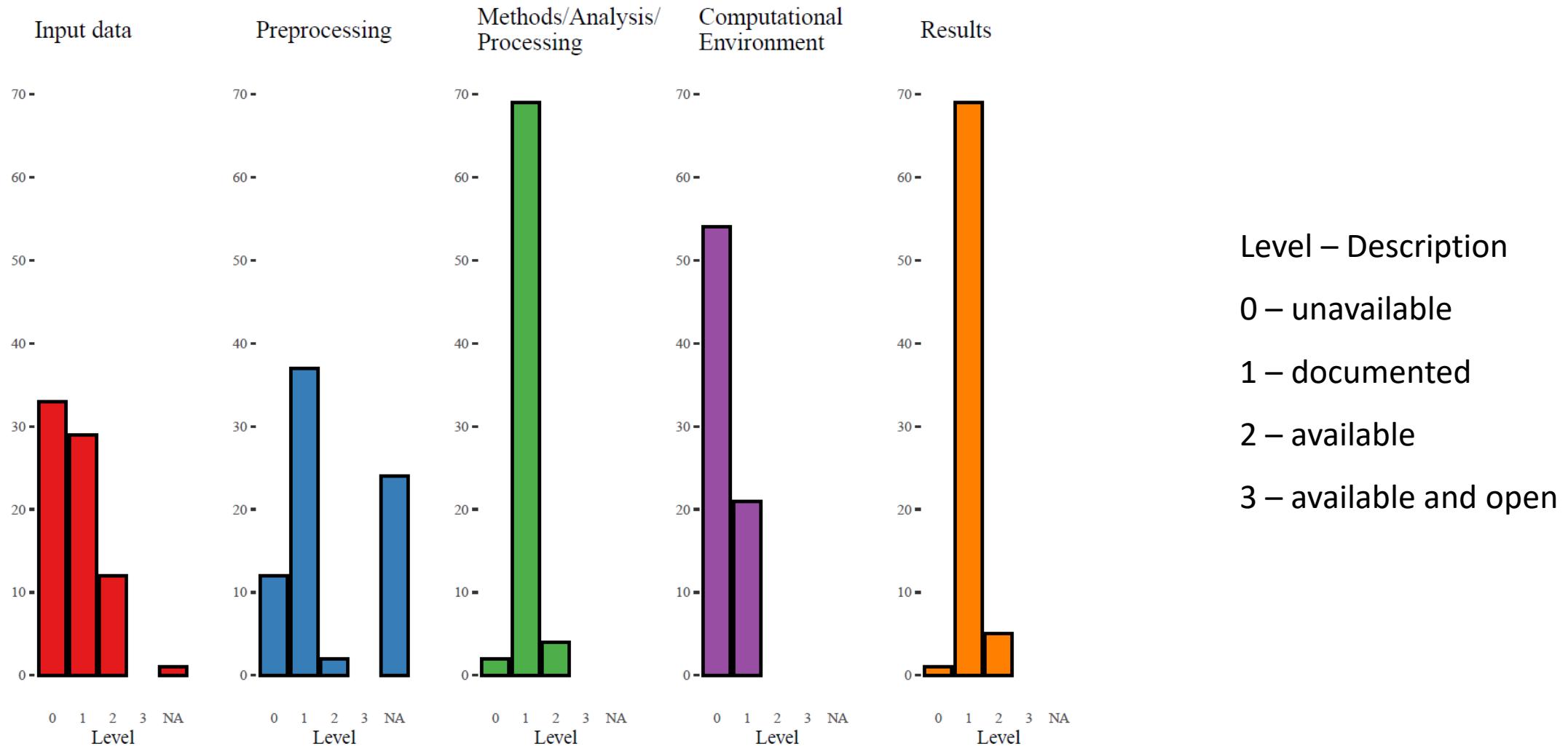
- labor-intensive, thus difficult to scale up
- *Preprocessing* not too helpful criterion (overlap with *Analysis*)
- *Computational environment* of limited use because relates mostly to processing time

Future replications should drop *preprocessing* and could drop *computational environment* criteria

Try it out!

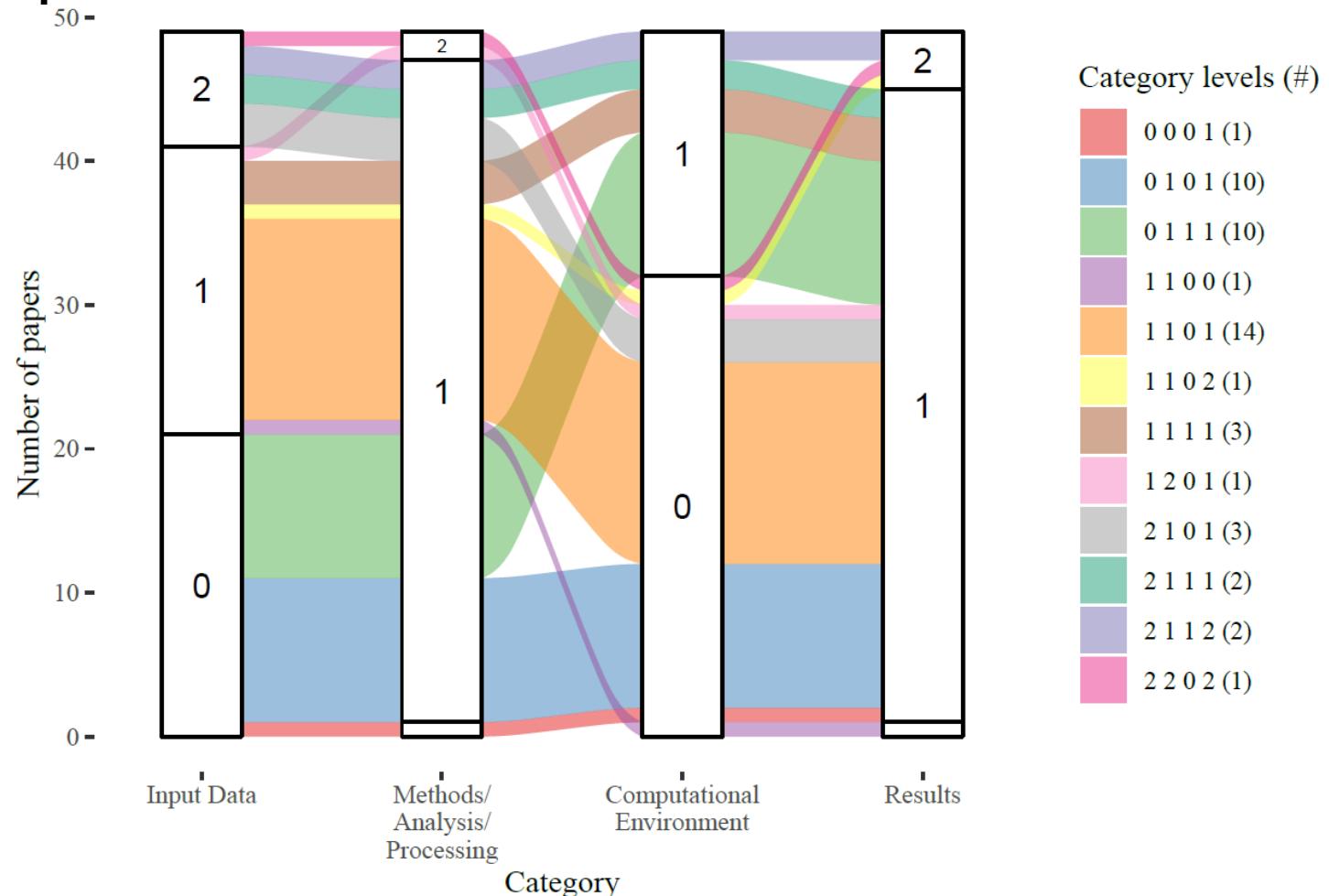
<https://github.com/nuest/reproducible-research-at-giscience>

# What's the outcome for GIScience?



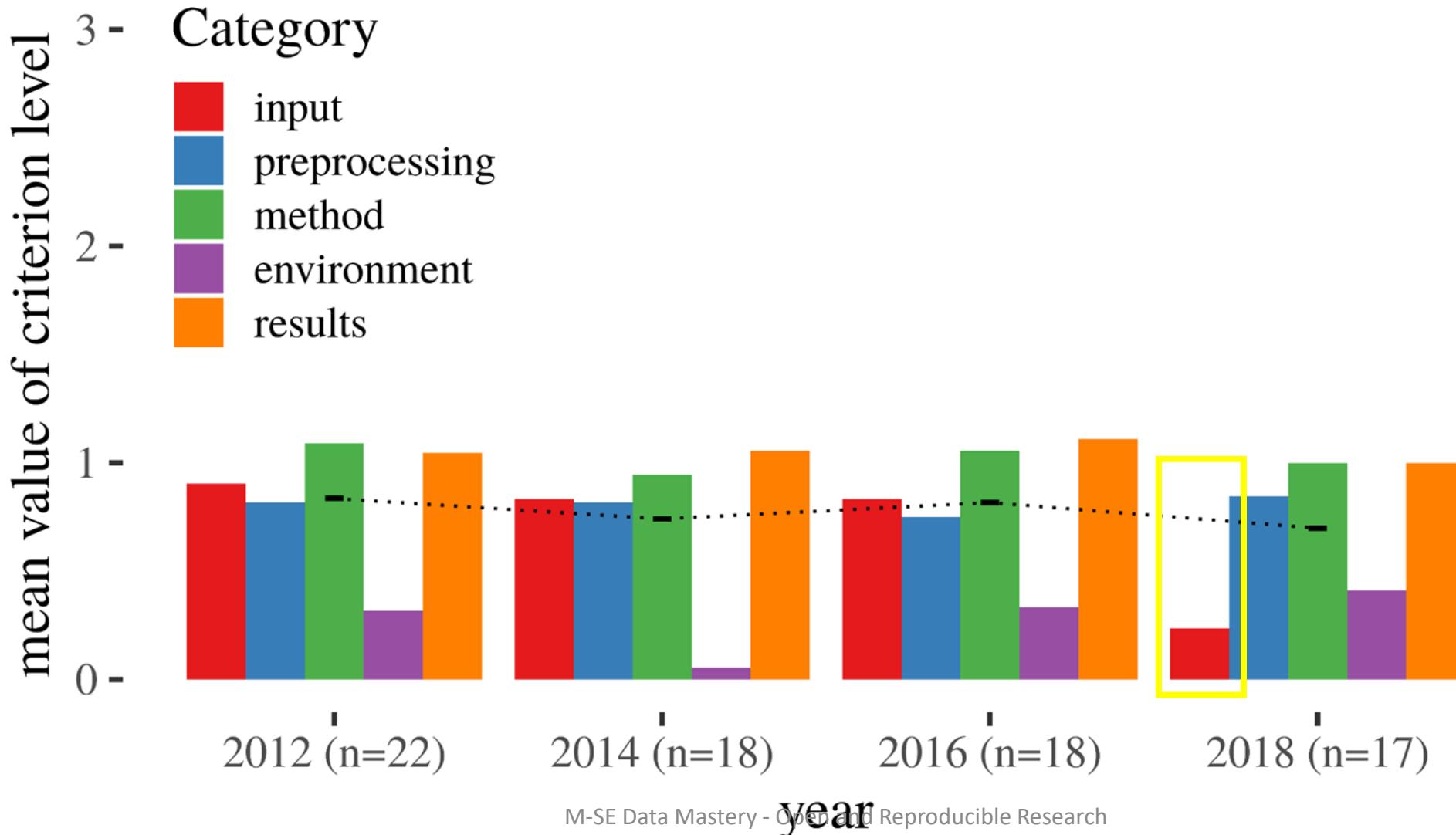
■ **Figure 1** Barplots of reproducibility assessment results; levels range from 0 (leftmost bar) to 'not applicable' (rightmost bar).

# Any patterns visible?



**Figure 2** Alluvial diagram of common groups of papers throughout 4 of 5 categories including only papers without any “not applicable” (*Level MA*) value; category *Preprocessing* was dropped because difficulty to clearly assess it lead to many “not applicable” values.

# Any change over time? (again, no)



# But what does this mean for GIScience?

- Overall reproducibility not great but: most papers meet standards for publication ('*documented*' in all three main criteria)
- Main problem is *input data* (several score only '*unavailable*')
  - Scores not a result of link rot (although that is a problem!): if there was reason to assume data was available at time of publication, paper received '*available*'
  - Worrisome, because of increased focus on data science and need for ML training data

# How do GIScience and AGILE compare?

■ **Table 3** Mean values per criterion for both conferences (rounded to two decimal places).

Criterion	AGILE full papers	GIScience papers
input data	0.67	0.72
method/analysis/processing	1.00	1.03
computational environment	0.62	0.28
results	0.88	1.05

- Similar in terms of topics
- overlap of authors noticeable but not majority
- different geographic scope
- Biannual vs annual
- AGILE has institutional framework (council) that supported newly implemented guidelines, reproducibility committee, and badges

# Perspective from another survey

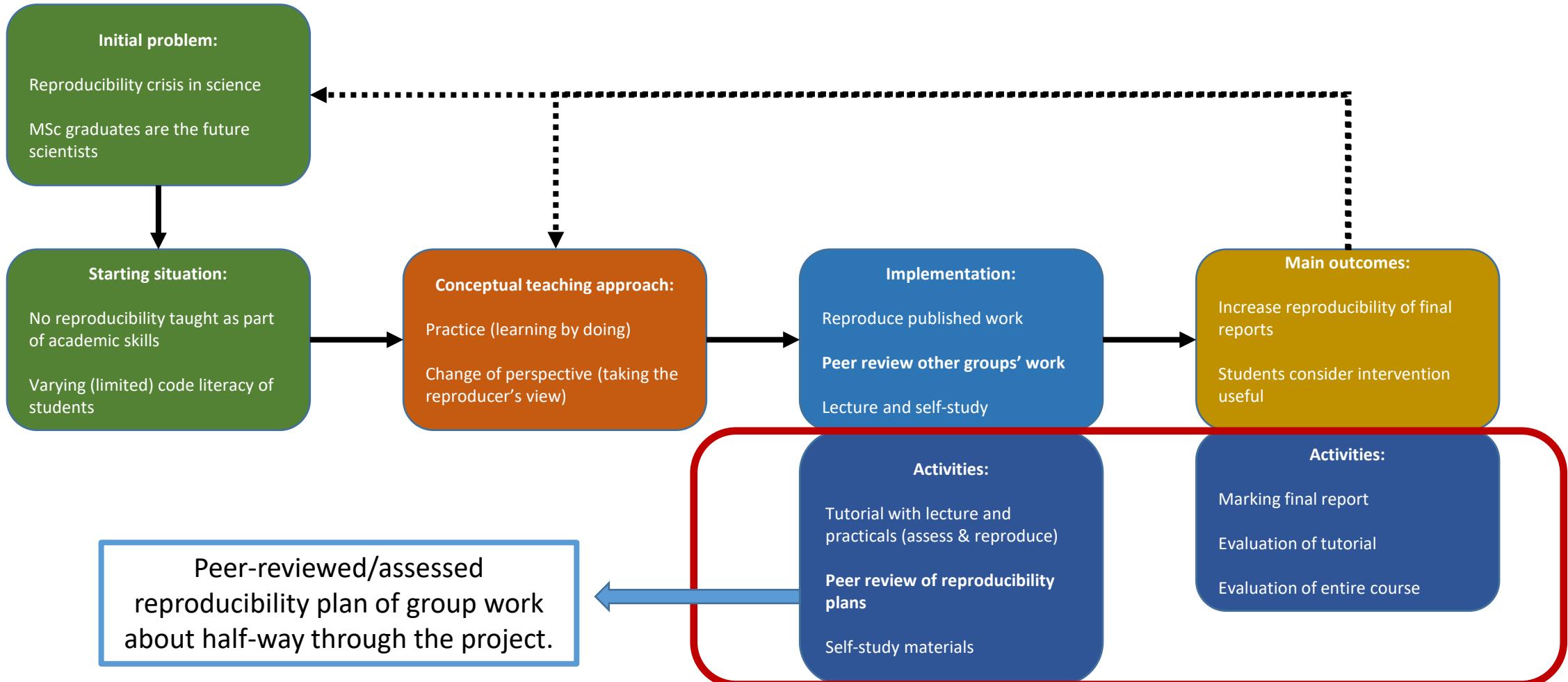
(Kedron et al. 2024, 10.1080/24694452.2023.2276115)

- Geographic researchers
    - Are aware of reproducible research practices
    - lack direct experience using those practices
  - Reproducibility is hindered by
    - Academic incentive systems
    - Inaccessibility of key components of prior research
- > Only a small percentage of researchers are attempting to independently reproduce past work.**

# Reproducibility in the classroom (here!)

- Data science and computational sciences demand algorithmic thinking and coding skills
- Open and reproducible research require specific skills for
  - Making data FAIR
  - Allowing replication and reproduction of publications
- In the geosciences, still a lot of focus on classic academic skills training and assessment:
  - Knowledge is tested in exams
  - Project work is not shared within course or beyond it
  - Process is less important than outcomes
  - Plagiarism is the ultimate sin, so refrain from re-using other people's work

# An intervention in Spatial Engineering



# What did the students think?

How useful did you find the ...	Not useful	A bit useful	Quite useful	Very useful
... introductory lecture on reproducibility?	0	2	5	3
... reading the example paper and scoring it (first part of the exercise)?	0	0	7	3
... reproducing the example analysis (second part of the exercise)?	1	4	4	1
... information on reproducibility strategies and recommendations?	0	1	4	5
<b>... the peer-reviewed reproducibility plan?</b>	<b>0</b>	<b>2</b>	<b>7</b>	<b>4</b>
<b>Summary</b>	<b>1</b>	<b>9</b>	<b>27</b>	<b>16</b>

# Part 2: What you can do

Based on reproducibility guidelines developed at expert meeting in April 2019

<https://osf.io/phmce/> and input from Dr. Markus Konkol

Available also in Spanish, other languages work in progress

"FINAL".doc



FINAL.doc!



FINAL\_rev.2.doc



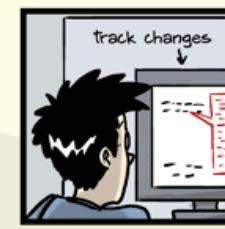
FINAL\_rev.2.doc



FINAL\_rev.6.COMMENTS.doc



FINAL\_rev.8.comments5.  
CORRECTIONS.doc



FINAL\_rev.18.comments7.  
corrections9.MORE.30.doc



FINAL\_rev.22.comments49.  
corrections.10.#@\$%WHYDID  
ICOMETOGRADSSCHOOL????.doc

JORGE CHAM © 2012

WWW.PHDCOMICS.COM

# How did the Reproducible AGILE Team form?

AGILE Conference **workshops**  
2017, 2018, and 2019

Review Paper 2017/2018

AGILE Initiative  
<https://o2r.info/reproducible-agile/2019/>

## Team

- [Daniel Nüst](#) (ifgi)
- [Carlos Granell](#) (Jaume I)
- [Barbara Hofer](#) (Z\_GIS)
- [Frank Ostermann](#) (ITC)
- [Rusne Sileryte](#) (TU Delft)

## Invited

## Experts

- [Anita Graser](#) (Austrian Institute of Technology)
- [Kristina Hettne](#) (CDS, Leiden University Library)
- [Karl Broman](#) (University of Wisconsin–Madison)
- [Marta Teperek](#) (TU Delft Library)



# The guidelines

<https://doi.org/10.17605/OSF.IO/CB7Z8>

## Reproducibility checklist

### Author guidelines

Writing DASA section

Data in Research Papers

Computational workflows in Research Papers

### Reviewer guidelines

### Reproducibility reviewer guidelines

### Background

## REPRODUCIBLE PAPER GUIDELINES

Full and short papers submitted to the AGILE conference **have** to include a **Data and Software Availability** section which documents data, software, and computational infrastructure to support reproduction, or mentions reasons for not publishing them.

The above requirement is the only one to comply with the AGILE Reproducible Paper Guidelines. The remainder of the document provides concrete recommendations for all involved stakeholders to increase transparency, reproducibility, and openness of computational GIScience research. The following table of contents shows the recommended parts for different readers. Familiarity with all sections is, of course, beneficial.



	<b>Reproducibility Checklist</b> Helps to ensure authors and reviewers do not miss anything important.	2
	<b>Author Guidelines</b> Show how to write the Data and Software Availability Section and give practical recommendations to make data and computational workflows reproducible.	4
	Writing the Data and Software Availability Section Including Data in Research Papers Including Computational Workflows in Research Papers	
	<b>Scientific Reviewer Guidelines</b> Describe role in evaluating plausibility and completeness of the data and software availability documentation.	7
	<b>Reproducibility Reviewer Guidelines</b> Describe role and approach to execute workflows and clarify efforts.	8
	<b>Background</b>	10

#### Further resources

These guidelines can not cover all details of the reproducibility review at AGILE conferences. For more information for authors, translations, and practical examples see the [guidelines wiki](#). For more information about the review process and deadlines, see the [process description](#). For any questions, please visit the AGILE Discourse server's [forum for the Reproducible Paper Guidelines](#).

# AGILE reproducible paper guidelines: Data

- ... has a clear (and permissive) license
- ... is accessible in a public repository
- ... is documented (at least description of collection query and field or column names; ideally complete metadata following established standards)
- ... is provided in a non-proprietary, text-based (human-readable) format, e.g., CSV

Legally open

Technically open

# AGILE reproducible paper guidelines: Analysis

- Computational environment (including hardware) is at least documented, or even provided (e.g., Docker container)
- The versions of relevant software components (libraries, packages) are provided
- Software is available in a public repository
- Software has a clear (permissive) license
- Computational steps are explained in a text file, flowchart, or script
- All parameters needed to run the computational workflow are provided
- All computations are mentioned

# AGILE reproducible paper guidelines: Output

- Datasets and code (including your own datasets and code used in the paper) are cited in the references with DOIs
- The reasons, if any, for not being able to share (parts of) data or code are clearly indicated
- All choices (filtering, design) made to create figures are explained

# What can you do today?

- Descriptive and consistent file and variable names  
([http://www2.stat.duke.edu/~rcs46/lectures\\_2015/01-markdown-git/slides/naming-slides/naming-slides.pdf](http://www2.stat.duke.edu/~rcs46/lectures_2015/01-markdown-git/slides/naming-slides/naming-slides.pdf))
- Don't modify the original data, ever! Try folder structures like:
  - Input: Original data, read-only! No preprocessing or intermediary results either.
  - Output: All files generated by your code, including preprocessing and intermediary results
  - Code: All source code and libraries
  - Documentation: All code documentation and manuscripts
- Plain text + git ([https://karthik.github.io/git\\_intro/#/five\\_min\\_break](https://karthik.github.io/git_intro/#/five_min_break))

# FAIR PRINCIPLES

## FAIR Principles

GO FAIR is committed to making data and services **findable, accessible, interoperable and reusable (FAIR)**.



**Findable:** Metadata and data should be easy to find for both humans and computers.



**Accessible:** The exact conditions under which the data is accessible should be provided in such a way that humans and machines can understand them.



**Interoperable:** The (meta)data should be based on standardized vocabularies, ontologies, thesauri etc. so that it integrates with existing applications or workflows.



**Reusable:** Metadata and data should be well-described so that they can be replicated and/or combined in different research settings.

(<https://www.nature.com/articles/sdata201618> ,

<https://www.force11.org/fairprinciples>)

## A Large-scale Study about Quality and Reproducibility of Jupyter Notebooks

João Felipe Pimentel\*, Leonardo Murta\*, Vanessa Braganholo\*, and Juliana Freire†

\*Universidade Federal Fluminense

Niterói, Brazil

{jpimentel,leomurta,vanessa}@ic.uff.br

†New York University

New York, USA

juliana.freire@nyu.edu

*Abstract*—Jupyter Notebooks have been widely adopted by many different communities, both in science and industry. They support the creation of literate programming documents that combine code, text, and execution results with visualizations and all sorts of rich media. The self-documenting aspects and the ability to reproduce results have been touted as significant benefits of notebooks. At the same time, there has been growing criticism of their result bad practices studied 1.4 analysis of also proposed reproducibility research areas. *Index Terms*—

its library dependencies with associated versions, which can make it hard (or even impossible) to reproduce the notebook. These criticisms reinforce prior work which has emphasized the negative impact of the lack of best practices of Software Engineering in scientific computing software [9], regarding separation of concerns [10], tests [11], and maintenance [12].

books is of and the ability. wide books. impact wide, both rated ected ation ; the s. As tions sion ; and ings, it of

But: Pimentel et al. studied 1.4 millions of notebook (GitHub). Only

24.11% of them run without exceptions, and only 4.03% produced the same results".

discuss the threats to the validity of our study in Section VI and present related work in Section VII. Finally, we conclude in Section VIII where we outline directions for future work.

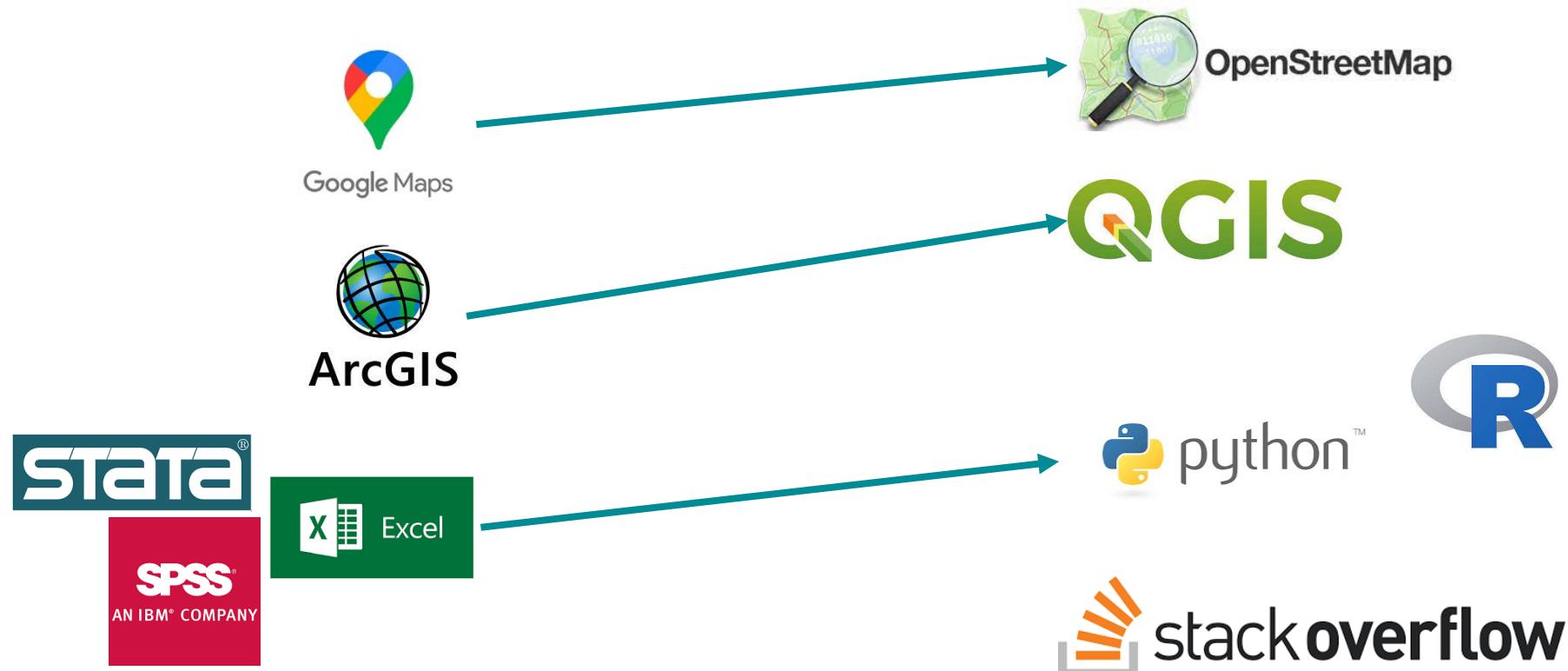
### II. BACKGROUND

Knuth [1] introduced the *literate programming* paradigm that, by combining code and natural language, enables programmers to explicitly state the thoughts behind a program's

# Five recommendations for ORR

---

**Recommendation 1:** Use open source software instead of commercial software.



# Five recommendations for ORR

---

## Recommendation 2: Learn a scripting language.



- Scripts describe every step of an analysis
- Human-readable description of what the code does
- Others can understand
  - What has been done
  - How it has been done
- Not reproducible
- No step-by-step description
- No control over the algorithms



# Five recommendations for ORR

---

**Recommendation 3:** Learn a computational notebook format.



# Five recommendations for ORR

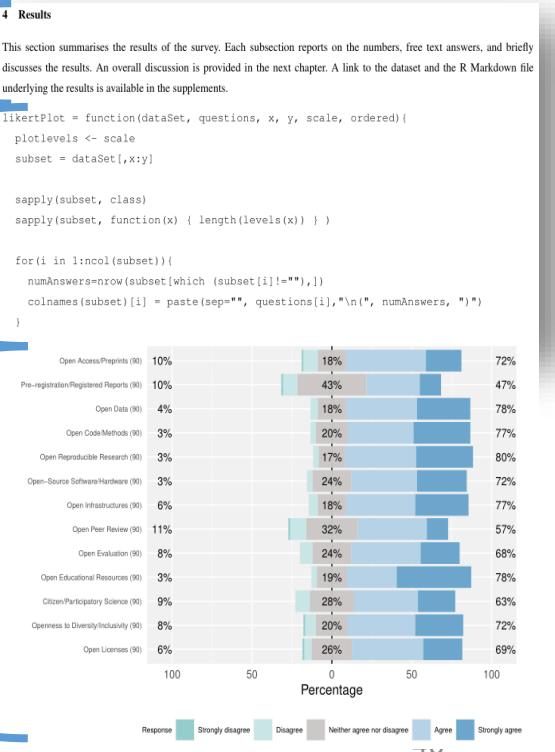
## Recommendation 3: Learn a computational notebook format.

```
138 # Results  
139  
140 This section summarises the results of the survey.  
141 Each subsection reports on the numbers, free text answers, and briefly discusses the results.  
142 An overall discussion is provided in the next chapter.  
143 A link to the dataset and the R Markdown file underlying the results is available in the supplements.  
144  
145 ````{r, echo=FALSE, results="hide", message=FALSE, comment=FALSE, warning=FALSE}  
146 likertPlot = function(dataSet, questions, x, y, scale, ordered){  
147   plotlevels <- scale  
148   subset = dataSet[,x:y]  
149  
150   sapply(subset, class)  
151   sapply(subset, function(x) { length(levels(x)) } )  
152  
153   for(i in 1:ncol(subset)){  
154     numAnswers=nrow(subset[which (subset[i]!=""),])  
155     colnames(subset)[i] = paste(sep="", questions[i], "\n(", numAnswers, ")")  
156   }  
157  
158   for(i in seq_along(subset)) {  
159     subset[,i] <- factor(subset[,i], levels=plotlevels)  
160   }  
161 }
```

Text

Code

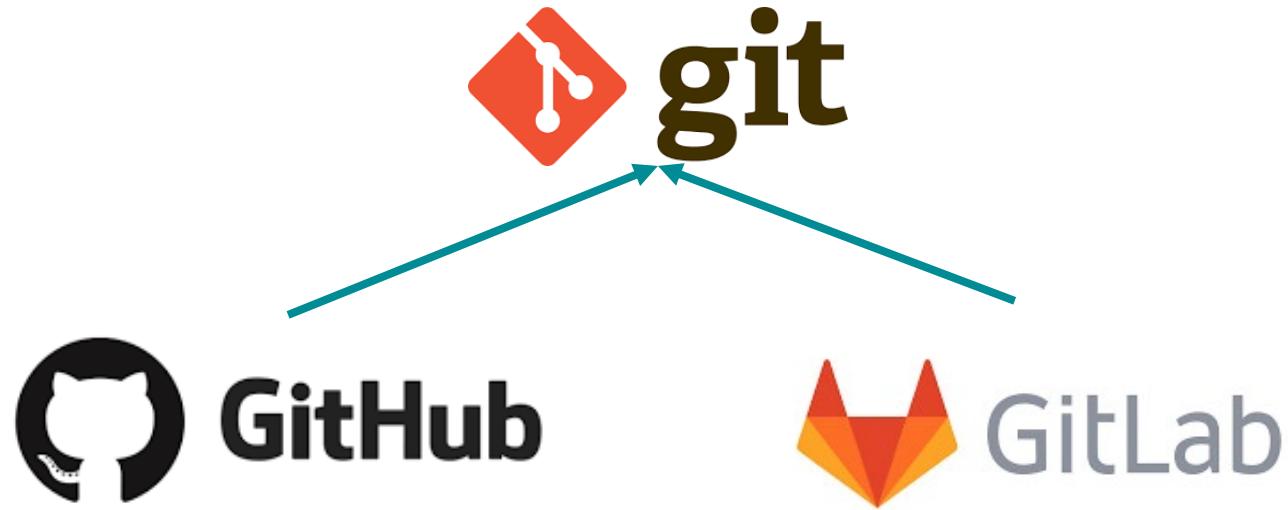
Output



# Five recommendations for ORR

---

**Recommendation 4:** Learn a collaborative software development tool.



# Five recommendations for ORR

---

## Recommendation 5: Document your source code.

- Create a clean workspace with a hierarchical folder structure and name files properly.
- Include a README text file to explain the code.
  - What does the software?
  - How can I install it?
  - Are there any computational requirements (e.g., operating system)?
  - How can I use it?
  - How long does the analysis take?
- Add a LICENSE, e.g., MIT License, APACHE License, or GNU.

	code	File folder
	data	File folder
	figures	File folder
	analysis	Firefox HTML Document
	analysis	RMD File
	LICENSE	Text Document
	README	Text Document

# Five recommendations for ORR

---

## Recommendation 5: Document your source code.

- **Findable:** Deposit source code in a repository that provides DOIs and metadata.
- **Accessible:** Provide the opportunity to download the source code in bulk form.
- **Interoperable:** Use open source programming languages and software.
- **Reusable:** Release the software under a clear and open usage license.

# Some useful platforms

OpenScienceFoundation (<https://osf.io/>) has many options, including anonymous sharing links

Zenodo (<https://zenodo.org/>) especially useful for preprints and data that need DOI

GitHub (<https://github.com/>) or GitLab (<https://gitlab.com/>)

Binder (<https://mybinder.org/>) for online executable notebooks

Google Colab for shared executable notebooks (<https://colab.research.google.com/> )

Re3Data.org (<http://re3data.org/>) registry of research data

# But I've completely ignored qualitative research?!?

- So qualitative research is not good science, because much of it is irreproducible?
- Of course not! I've done qualitative research myself and I know how valuable and difficult it is.
- Remember: Reproducibility is a spectrum. Let's try to make qualitative research as reproducible as possible!

# Qualitative research – some considerations

- Privacy often even more of a concern, because more in-depth information of individual participants is collected
- Many steps are inherently irreproducible (participants, setting, etc.)
- BUT: replication is certainly an option!
- Provide
  - As much information (demographics) on participants as possible
  - Share anonymized transcripts
  - Maybe document laboratory setting with a video

Most importantly:

**Some reproducibility is always  
better than no reproducibility!**

If the gold standard seems unreachable,  
don't give up, do what you can!

## FOR NEXT WEEK

- Read "[Why Data Is Never Raw](#)" and *think* about it! (i.e., don't outsource it to some artificial "intelligence", a.k.a. stochastic parrot)
- Come up with three reasons why Open Science is a *bad* idea!
- Best approach:
  - Do this *right after this class* and take *notes*
  - *Read them again* next Wednesday morning before class