Reproducible Research in the **Quantitative** Geo**Information** Sciences

Frank Ostermann (UT, presenter)
based on work with
Carlos Granell (UJI), Barbara Hofer (PLUS),
Markus Konkol (WWU/UT), Daniel Nüst (WWU), Rusne Sileryte (TUD)

Home About

Search Twitter

Have an account? Log In

Sylvain ❄️🎓

@DevilleSy

Freezing stuff since 1876. Will science for chocolate. ORCID Id 0000-0002-3363-3184. Author of "Freezing Colloids" springer.com/fr/book/978331…

France

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

54 2.6K 5.9K

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

Why Most Published Research Findings Are False

John P. A. Ioannidis

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

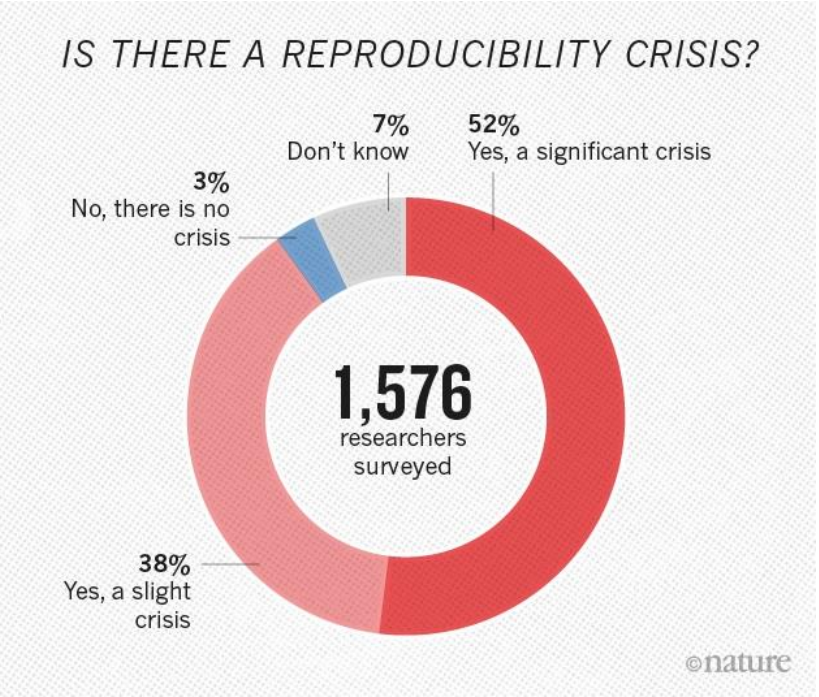
https://www.vox.com/future-perfect/21504366/science-replication-crisis-peer-review-statistics

Published: 25 May 2016

1,500 scientists lift the lid on reproducibility

Monya Baker

Nature **533**, 452–454 (2016)



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

Artificial intelligence / Machine learning

AI is wrestling with a replication crisis

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

by **Will Douglas Heaven**
November 12, 2020

https://www.technologyreview.com/2020/11/12/1011944/artificial-intelligence-replication-crisis-science-big-tech-google-deepmind-facebook-openai/

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)

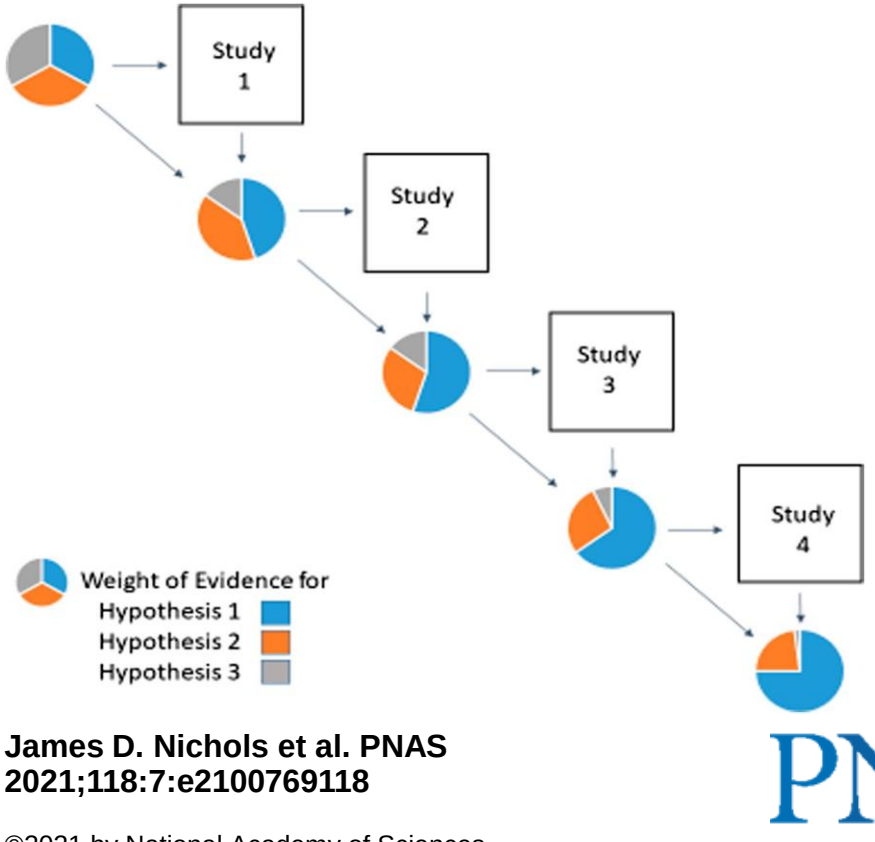
If outcomes are similar, original hypothesis is supported

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

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?)



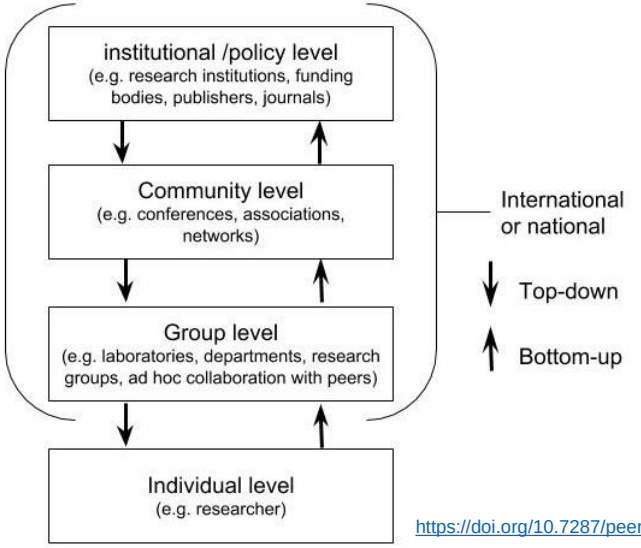
James D. Nichols et al. PNAS 2021;118:7:e2100769118

PNAS

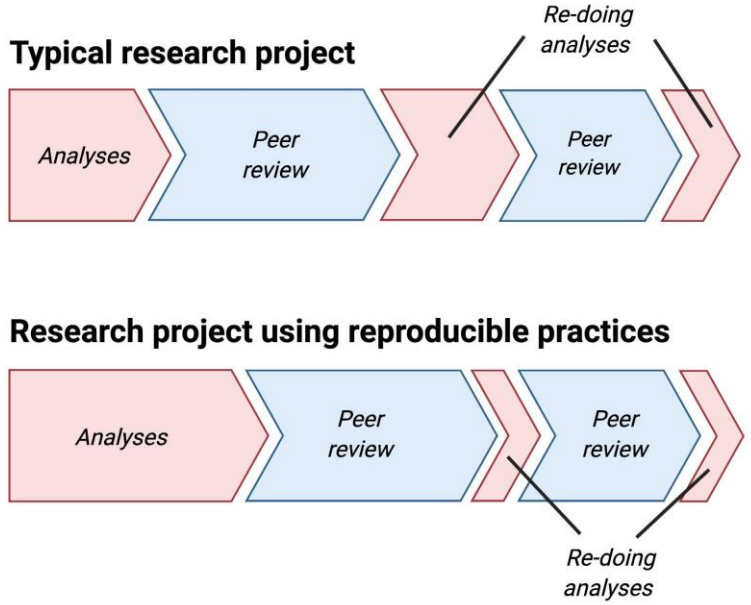
Why do they matter?

For individual actors:

- 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

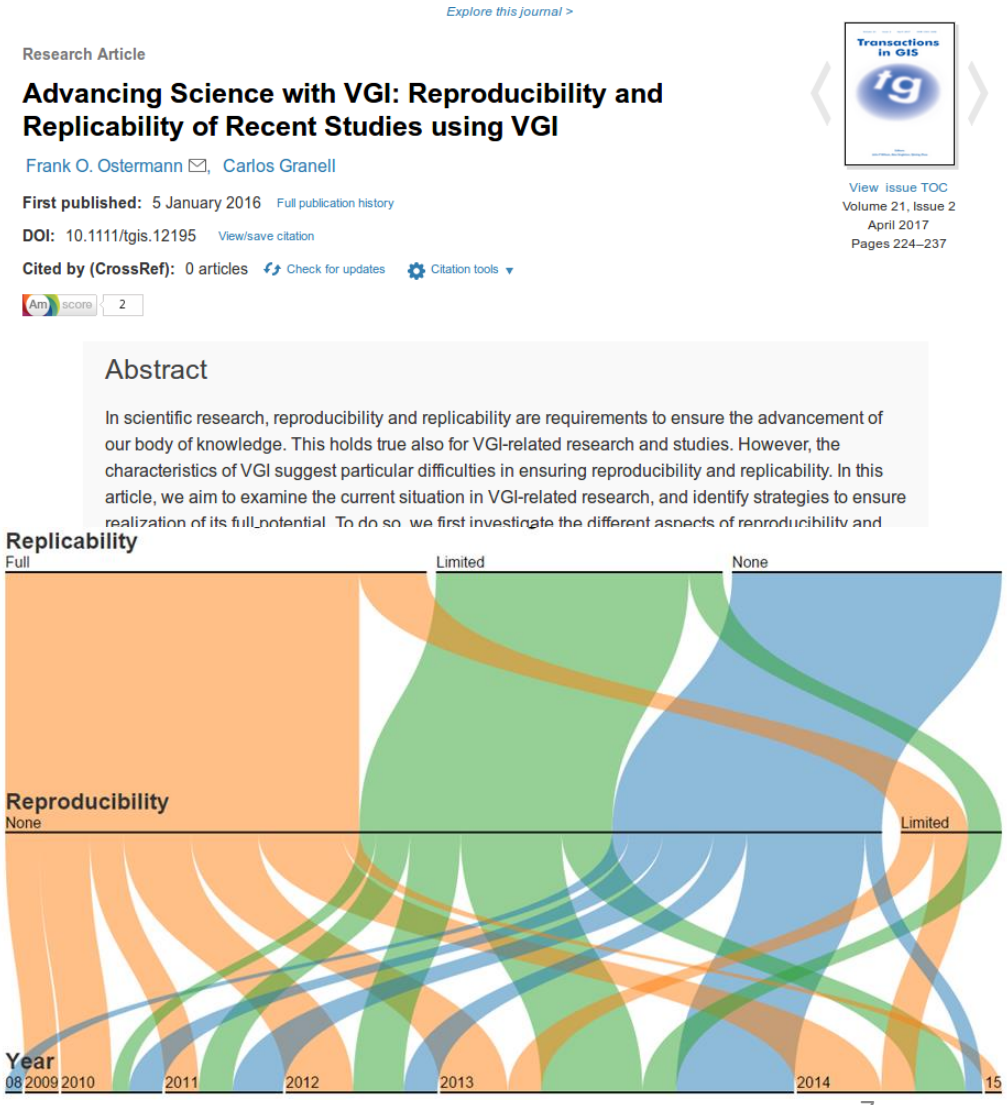


Quintana, D. S. (2020, November 28). Five things about open and reproducible science that every early career researcher should know. https://doi.org/10.17605/OSF.IO/DZTVQ

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

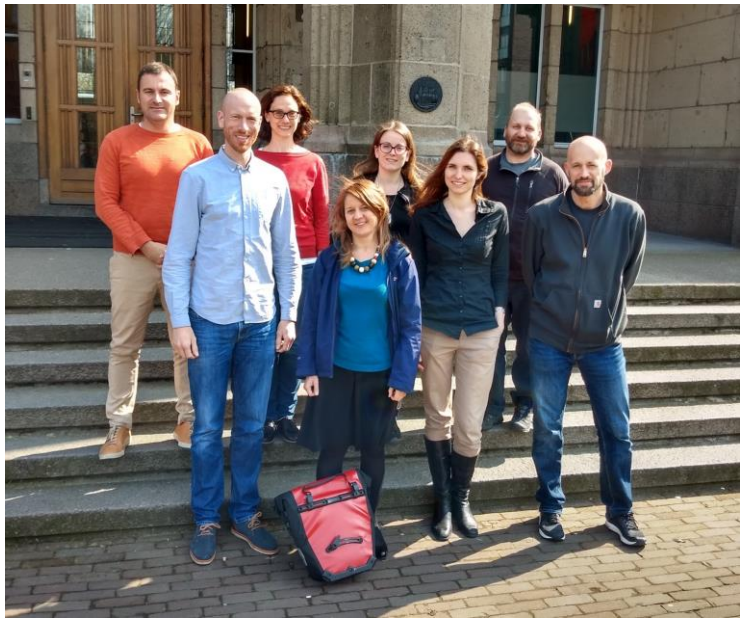


How did the Reproducible AGILE Team form?

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

Review **Paper** 2017/2018

- Daniel Nüst (ifgi)
- Carlos Granell (Jaume I)
- Barbara Hofer (Z_GIS)
- Frank Ostermann (ITC)
- Rusne Sileryte (TU Delft)



AGILE **Initiative**

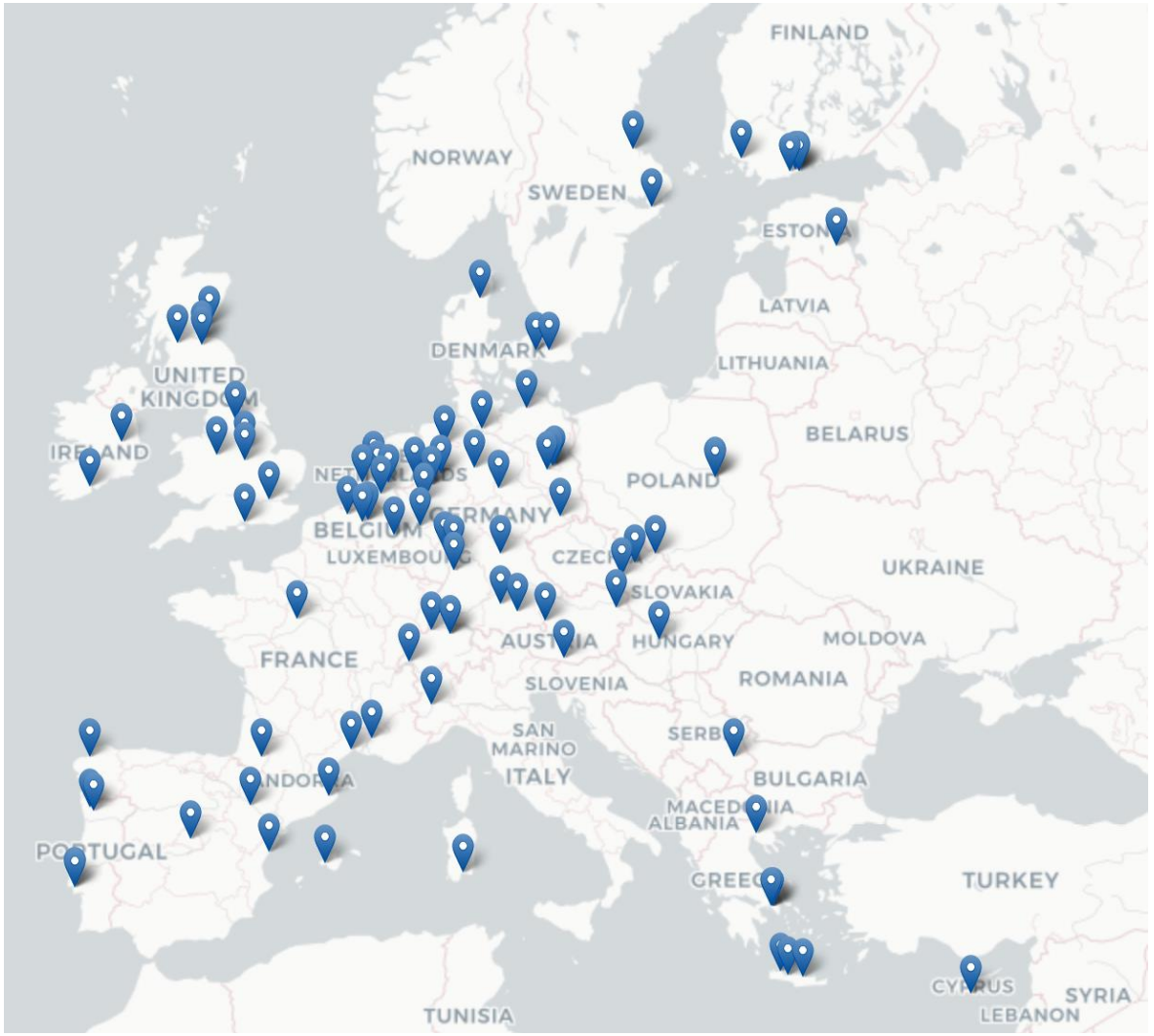
https://o2r.info/reproducible-agile/2019/

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

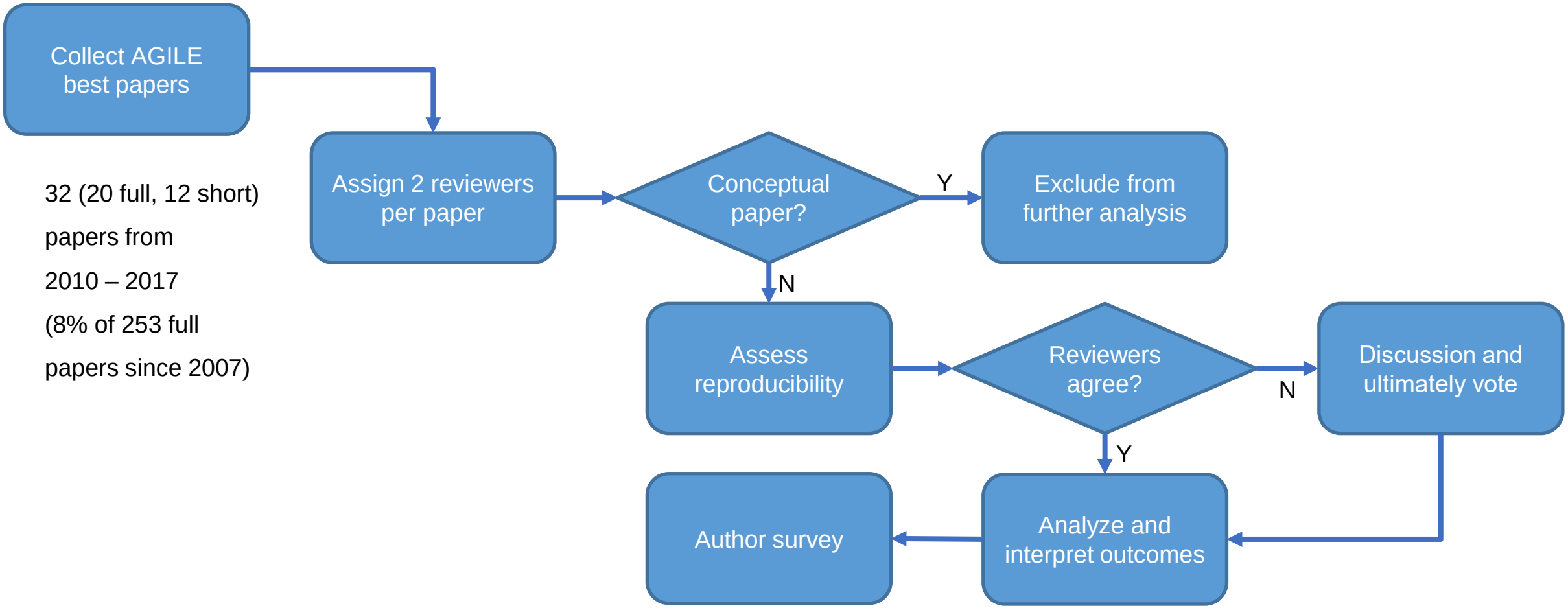
Wait, what's AGILE?

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

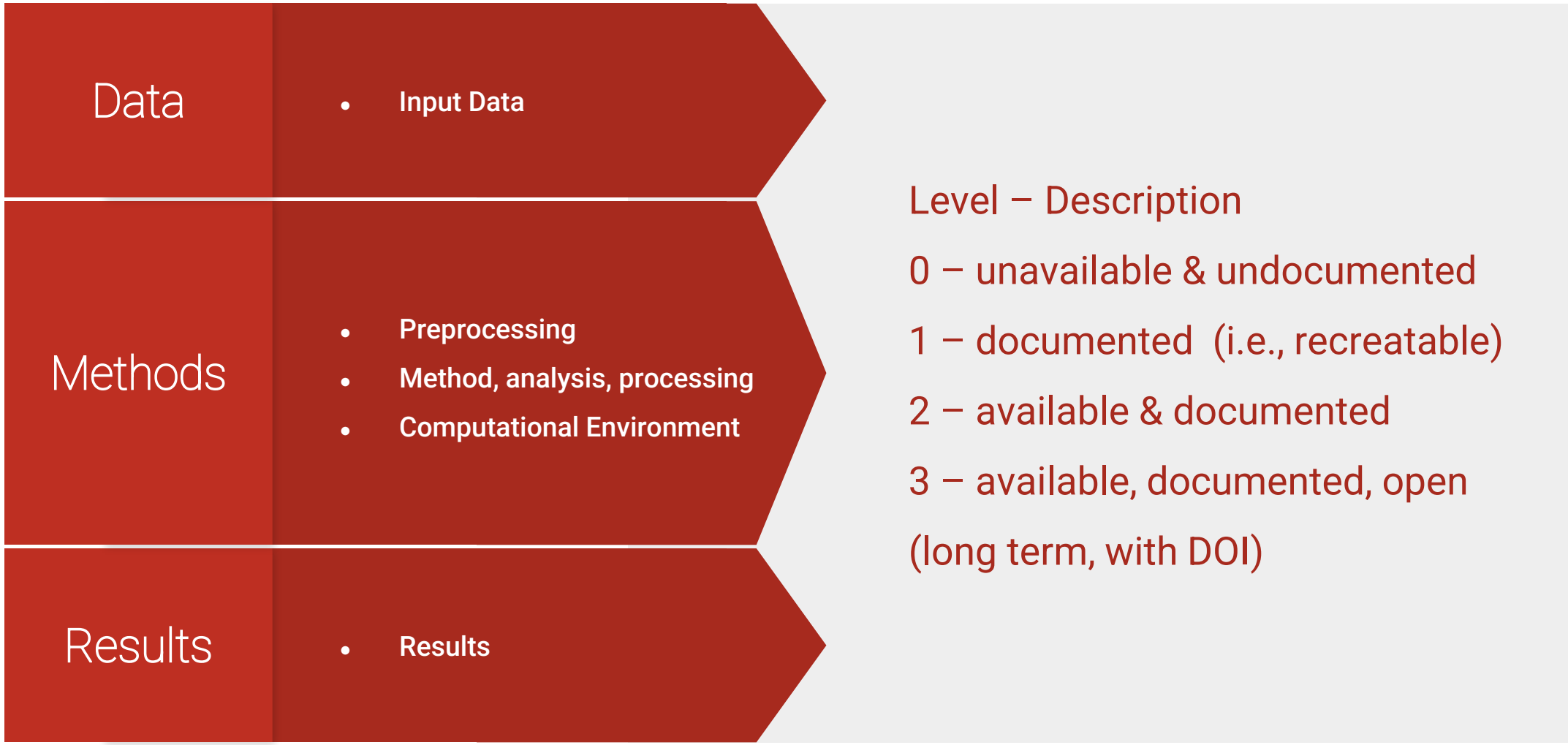
- Annual AGILE conference (https://agile-online.org/conference-2020)
- Bi-annual PhD School (https://agile-online.org/agile-actions/phd-school)
- AGILE Initiatives (https://agile-online.org/funding-initiatives)
- Collaboration & MoU with organizations & sister associations (https://agile-online.org/agile-community/cooperation)



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

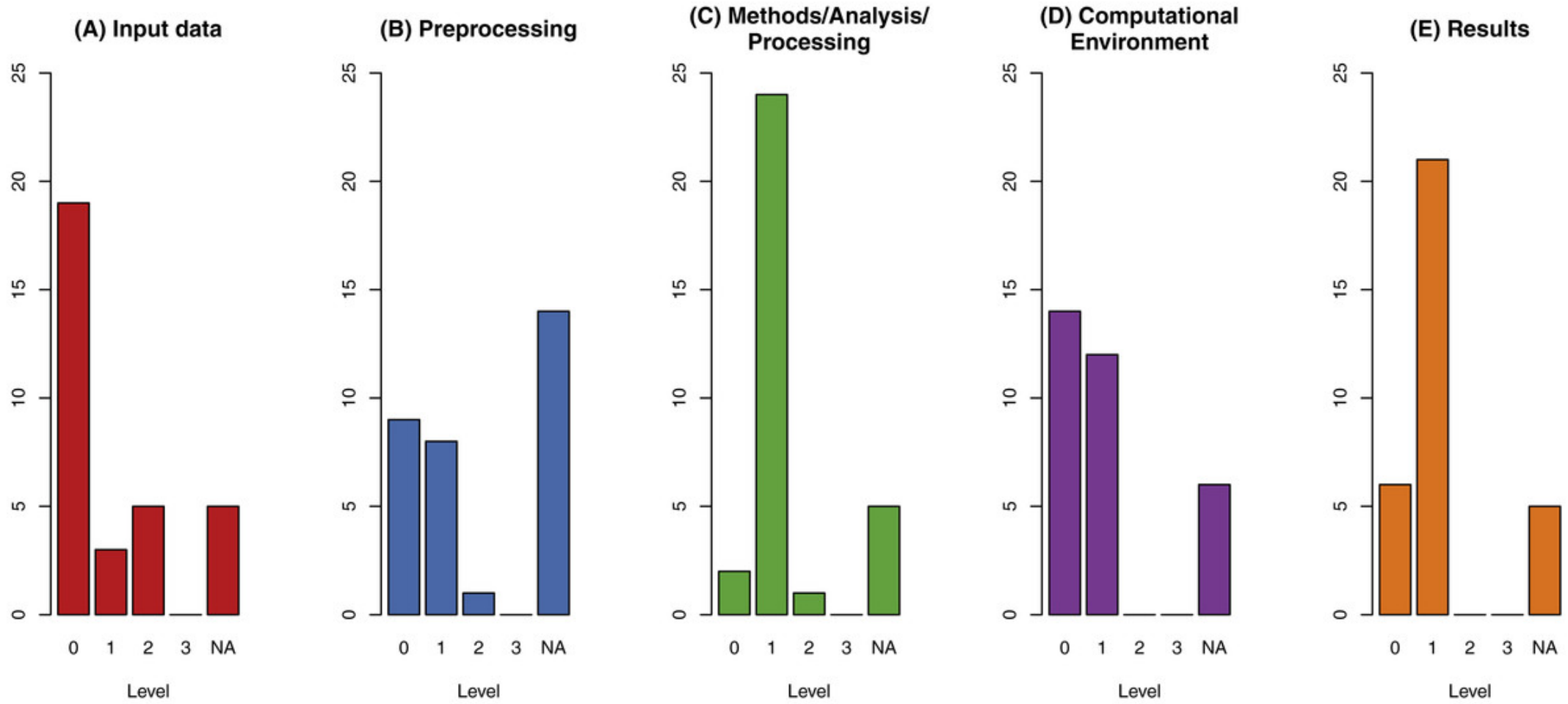


How can one assess reproducibility?

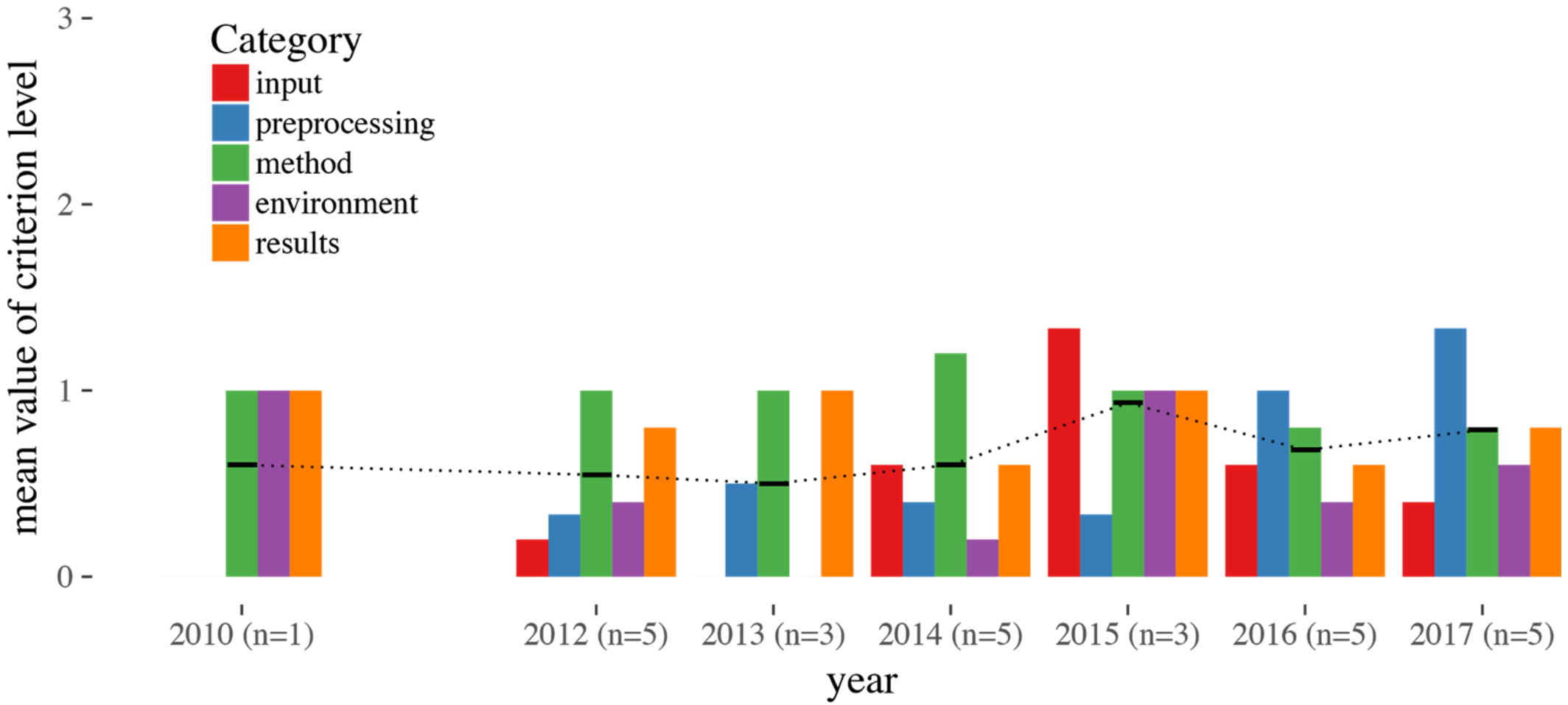


Nüst D, Granell C, Hofer B, Konkol M, Ostermann FO, Sileryte R, Cerutti V. (2018)
Reproducible research and GIScience: an evaluation using AGILE conference papers.
PeerJ 6:e5072 https://peerj.com/articles/5072

How reproducible were AGILE papers?



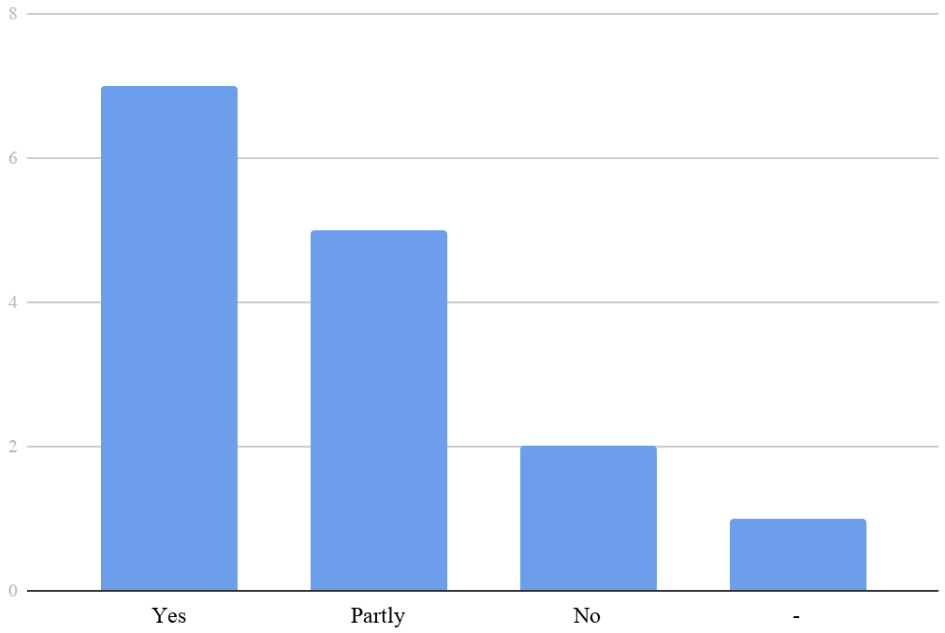
Does it at least improve over time? (no)



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?

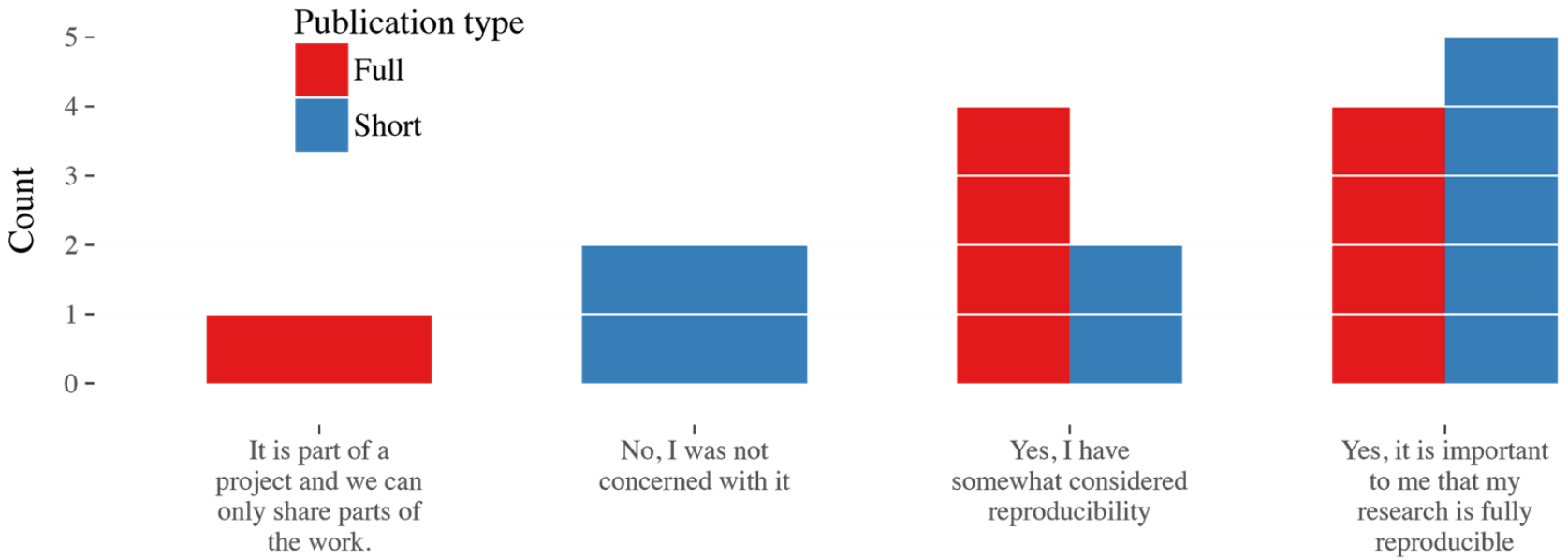


Reasons for disagreement:

- 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

AGILE Reproducible Paper Guidelines: Contents, first revision (2020), and outcomes of 2021 reviews

(slides by Daniel Nüst, modified by FO)

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

Website: https://osf.io/phmce/
Version: December 2020
DOI: 10.17605/OSF.IO/CB7Z8

AGILE Association of Geographic Information Laboratories in Europe

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.

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.

The guidelines for reproducibility reviewers

Ideal vs. realistic

Role & skills

- Do shift burden to author
- Do encourage and set examples
- Do *not* accept private data sharing
- Document your work in report (impact)
- Be kind (career stage, knowledge, privileges)

REPRODUCIBILITY REVIEWER GUIDELINES

Reproducibility reviewers conduct a complimentary review of the computational workflow that is published with a full paper that is provisionally accepted after the scientific review process. They read the paper insofar as needed to **reproduce the computation**, **using the abstract and the Data and Software Availability section** (DASA) as starting points. Ideally, these sections of the paper together with a README file are sufficient for the reproduction. When reproducibility reviewers get stuck, they take advantage of the option to **communicate** with the authors early and often. Reproducibility reviewers should be aware of the different reproducibility levels (see Author Guidelines above) to **recommend improvements** to the authors, but they are not responsible for making a workflow transparent or executable. Reproducibility reviewers **write a reproducibility report** documenting the results of their reproduction attempt and their communication with the authors. The report is published if the reproduction was, at least in part, successful. It is shared with the authors if the reproduction attempt was stopped but already contains relevant feedback.

Reproducibility review coordination

The reproducibility chair will be your contact person regarding supporting infrastructure and getting access to the private discussion forum for reproducibility reviewers on the AGILE Discourse server[29]. This forum is used to assign, under the leadership of the reproducibility chai
respective topical and technical skills, and share mat
report.

Goals and scope

While the AGILE reproducible paper guidelines are
reproducibility success rate for accepted papers,
understanding, and ultimately community adoption thr
tasks as reproducibility reviewer harder and progress
review is an extra merit for an accepted paper, bu
acceptance. The reproducibility reviewer should be awa
might "take the extra few steps" needed. This non-exc
one reproducibility reviewer is assigned per paper. Y
scientific reviewer on the same paper, but the roles of th
of the reproducibility review is roughly in line with t
community is worth exploring for further examples and
reproduction, e.g., the recreation of some but not all of t
though what is "good enough" may change over time.
or the reproducibility committee chair in case of doubt.

Reproducibility reviewer skills

A reproducibility review is a learning experience for bo
AGILE community to increase openness and transparer
amount of time you should spend on a reproduction at
as the research you are tasked to reproduce. However
few minutes of being stuck and not spending more t
depends also on your interest, time budget, and skills
get basic familiarity with package managers and virtua
DESCRIPTION files and renv for R, npm for JavaSc
reproducibility reviewer discussion forum early and often

Do	Don't
Quick pre-repro-review checks and ask authors to fix before continuing; even if not all of these are technically required, authors who are willing to work reproducibly can show their engagement right from the start: 1. Do the links to data sets and materials resolve? 2. Is there a README with clear step-by-step instructions? 3. Is there a clear mention of to be expected execution times? 4. Is there a LICENSE file to ensure openness?	Dig across badly or un-documented collections of files and functions to identify which part of the code/data creates which figure/table/output; find or build the "start button" yourself.
Encourage authors by pointing out promising intermediate results or concrete benefits of reproducibility.	Run workflows requiring considerable computational resources (unless interesting for you) but ask for data subsets for demonstration purposes.
Accept sample datasets to run a workflow and compare the outcome with the expected sample results; check the sources of the full datasets, if available.	Accept private sharing of data or code, unless strictly required for protection of sensitive data. All changes by the author should update to the public reproduction material.
Clearly document the extent of the reproduction in your reproduction report and suggest potential improvements; if you provide intermediate feedback, to include a history of your interactions in the report so that the ideas you contributed are preserved when the submission's material is improved.	Attempt to install software without any instructions, install binary software of unknown origin, or try to fix installation problems you encounter on your machine; try to install without (a) asking for help from a fellow reproducibility reviewer who is familiar with the software, or (b) asking the author to help, providing a minimal reproducible example of your problem.
Get in touch with fellow reproducibility reviewers if specific expertise (tool, programming language, ..) is needed.	Point out or even fix problems that are not specific to the submission, e.g., general problems in a software tool.
Set an example when communicating about computational problems, e.g., by clearly defining your system (OS version, language version, etc.)	Create accounts on any service or platform to access code, data, or other resources.
Ask specific questions or point out concrete problems that may lead authors to improve their material, including referencing these guidelines or concrete tools/methods that you already (!) know about, especially if you suspect that the author might now be familiar with them (e.g., version pinning/dependency management, absolute paths).	Fix anything (unless you really enjoy doing so), e.g., • compiler problems, • outdated libraries, • broken paths, or • Incomplete computing environment specifications, especially if the author can fix them even quicker.
Make sure that you are aware of any templates or specific resources provided for reproducibility reviewers from the reproducibility committee chair before starting your review.	
Consider the author's background, career stage, and position to be aware of (a lack of) privileges or institutional power to decide how much support you provide and how you communicate; your reproducibility review can be a contribution to improve equity and inclusion in academia.	Be a bro.

Review process

Proceedings:
https://www.agile-giscience-series.net/review_process.html

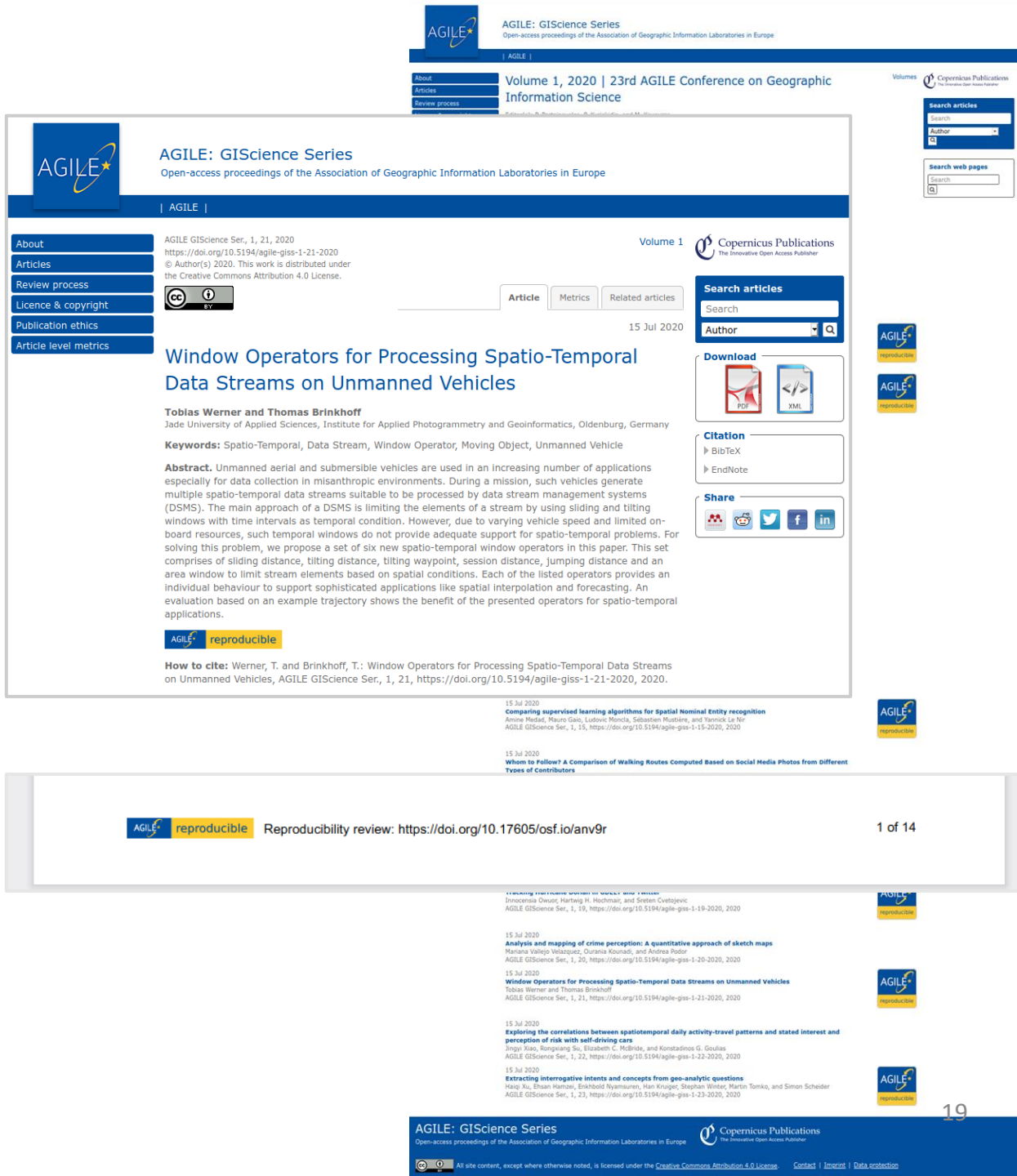
Process documentation:
https://osf.io/7rjpe/

Reproducibility review *after* accept/reject decisions

Reproducibility review & communication

Community conference & volunteers

Badges on proceedings website, article website with link, and first article page
(NEW! Thanks you, Copernicus!)



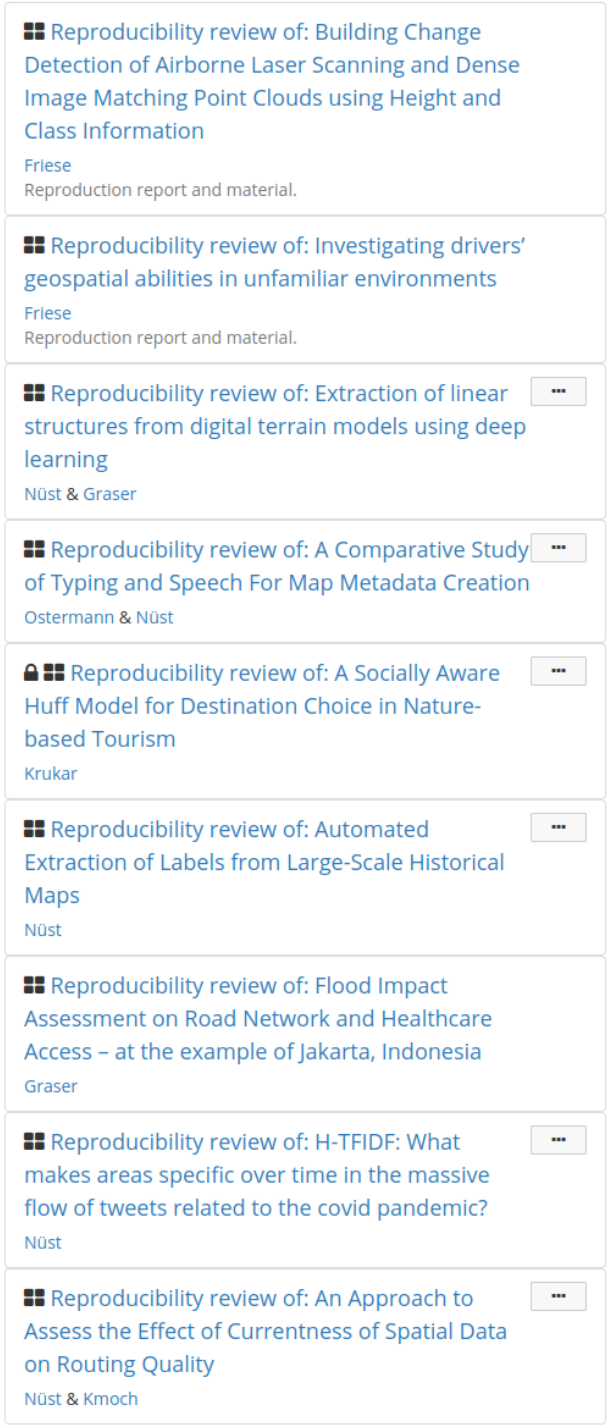
Reproducibility Review Outcomes

9 reproducibility reports published (2020: 6)

8 not reproducible:

- 3 conceptual papers
- **data not shared (choice, licence)**
  - synthetic data! subsets!
- **code** not shared (choice) or proprietary software

(repro reviewer matching failed)



https://osf.io/h64sd/

Reproducibility Reports

Published on OSF with a DOI

Title page, cites the paper

Paper links to report via URL

(no citation)

Automatically added to ORCID profile

Eventually indexed in GS

Reproducibility review of: Investigating drivers' geospatial abilities in unfamiliar environments

Philipp A. Friese

2021-06-07

REPRODUCIBLE AGILE

This report is part of the reproducibility review at the AGILE conference. For more information see https://reproducible-agile.github.io/. This document is published on OSF at https://osf.io/dx92a. To cite the report use

Friese, Philipp A. (2021, May). Reproducibility review of: Investigating drivers' geospatial abilities in unfamiliar environments. https://doi.org/10.17605/OSF.IO/DX92A

Reviewed paper

Karkasina, D., Kokla, M., and Tomai, E.: Investigating drivers' geospatial abilities in unfamiliar environments, AGILE GIScience Ser., 2, 3, https://doi.org/10.5194/agile-giss-2-3-2021, 2021.

Summary

The updated submiss … t, dataset and questionnaires. The provi … paper and generates

2.4 Data and Software Availability

Questionnaires and sketches were collected anonymously. All statistical analyses, which results are detailed in the following section, have been performed in R (R Core Team, 2021) using the tidyverse package (Wickham et al., 2019). Driving directions given to participants, an Exemplary Questionnaire in English, the collected survey data in tabular form, the R code of the statistical analysis workflow, and all necessary metadata supporting this publication, are available on figshare and are accessible via the following DOI: https://doi.org/10.6084/m9.figshare.14460102.v4. The workflow underlying this paper was successfully reproduced by an independent reviewer during the AGILE reproducibility review and a reproducibility report was published at https://doi.org/10.17605/OSF.IO/DX92A.

ORCID Connecting Research and Researchers

ABOUT FOR RESEARCHERS MEMBERSHIP DOCUMENTATION

Daniel Nüst

Biography

Employment (6)
Education and qualifications (2)
Invited positions and distinctions (1)
Membership and service (5)
Funding (3)
Works (50 of 74)

Your new notifications

YOUR RECORD

DataCite has made changes to your ORCID record

Showing 5 out of 5 changes made by this client

WORKS

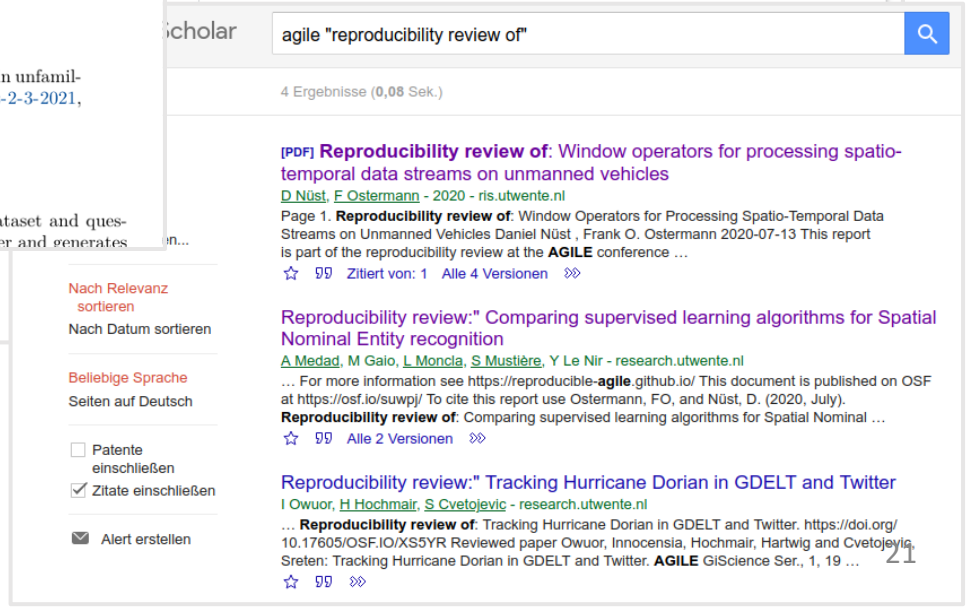
Added

- Reproducibility review of: A Comparative Study of Typing a (2021-06-08)
- Reproducibility review of: An Approach to Assess the Effec (2021-06-08)
- Reproducibility review of: Automated Extraction of Labels f
- Reproducibility review of: Extraction of linear structures fro (2021-06-08)
- Reproducibility review of: H-TFIDF: What makes areas spe to the covid pandemic? (2021-06-08)

Reproducibility review of: A Comparative Study of … Creation
Open Science Framework
2021 | other
DOI: 10.17605/osf.io/7fqtm
Source: DataCite

Reproducibility review of: An Approach to Assess … Data on Routing Quality
Open Science Framework
2021 | other
DOI: 10.17605/osf.io/bdu28
Source: DataCite ★ Preferred source

Reproducibility review of: Automated Extraction of Labels from Large-Scale Historical Maps
Open Science Framework
2021 | other
DOI: 10.17605/osf.io/anv9r



Scholar

agile "reproducibility review of"

4 Ergebnisse (0,08 Sek.)

[PDF] **Reproducibility review of**: Window operators for processing spatio-temporal data streams on unmanned vehicles
D Nüst, F Ostermann - 2020 - ris.utwente.nl
Page 1. **Reproducibility review of**: Window Operators for Processing Spatio-Temporal Data Streams on Unmanned Vehicles Daniel Nüst , Frank O. Ostermann 2020-07-13 This report is part of the reproducibility review at the **AGILE** conference …
Zitiert von: 1 Alle 4 Versionen

Reproducibility review:" Comparing supervised learning algorithms for Spatial Nominal Entity recognition
A Medad, M Gaio, L Moncla, S Mustière, Y Le Nir - research.utwente.nl
… For more information see https://reproducible-**agile**.github.io/ This document is published on OSF at https://osf.io/suwpj/ To cite this report use Ostermann, FO, and Nüst, D. (2020, July). **Reproducibility review of**: Comparing supervised learning algorithms for Spatial Nominal …
Alle 2 Versionen

Reproducibility review:" Tracking Hurricane Dorian in GDELT and Twitter
I Owuor, H Hochmair, S Cvetojevic - research.utwente.nl
… **Reproducibility review of**: Tracking Hurricane Dorian in GDELT and Twitter. https://doi.org/10.17605/OSF.IO/XS5YR Reviewed paper Owuor, Innocensia, Hochmair, Hartwig and Cvetojevic, Sreten: Tracking Hurricane Dorian in GDELT and Twitter. **AGILE** GiScience Ser., 1, 19 …

Nach Relevanz sortieren
Nach Datum sortieren
Beliebige Sprache
Seiten auf Deutsch
Patente einschließen
Zitate einschließen
Alert erstellen

General observations and lessons learned

- **Further improvement over last year:** better prepared workflows; remaining hurdles: insufficient documentation, no “quick” executaion variant or lack of expected data size/runtime, links Figures < > Scripts
- **Community understanding improving but still needs time**: Had to remind authors to add DASA section - how can we be clearer in the communication? Camera-ready papers by authors possible, but exhausting.
- **Additional reproducibility questions for scientific reviewers worked:** too many submission to check for repro chair
- **Repro reviews less strict than originally planned:** promote positive examples and don’t expect perfection
- **Non-blindness:** served its purpose but unblinding also delayed procedures
- **Schedule still a challenge:** partly because infrastructure (EasyChair) does not enable reviewer roles and communication; workarounds with scripts and scraping
- **Improvements to process**: clarity that **DASA section is mandatory**, do not offer authors to object to report publications (no problems!)
- **Reproduction not attempted != bad science:** reproducibility is a ***spectrum;*** continue education on reproducibility, increase requirements while practices spread in community

How to put your community on a path towards more reproducibility in 5 ~~easy~~ hard steps

1. Build a team of enthusiasts (workshop, social events)
2. Assess the current state and raise awareness (workshop, paper)
3. Institutional support (🙏 AGILE Council 🙏 + committee chairs)
4. Positive encouragement (no reproduction != bad science)
5. Keep at it!

Next steps

Do it again in 2022

Revise guidelines? IT FR CN

Grow reproducibility reviewer team

YOU!, opportunity ECRs (mentoring/workshops/…)

Continue meta-research

Ostermann, F., Nüst, D., Granell, C., Hofer, B., & Konkol, M. (2020). *Reproducible Research and GIScience: an evaluation using GIScience conference papers*. EarthArXiv. https://doi.org/10.31223/x5zk5v

Continue community engagement towards opening scholarship

Scope

Requirements

Acceptance condition?

Open review if tenured

Format-free first submission

CRediT

Phase out when standard practice...

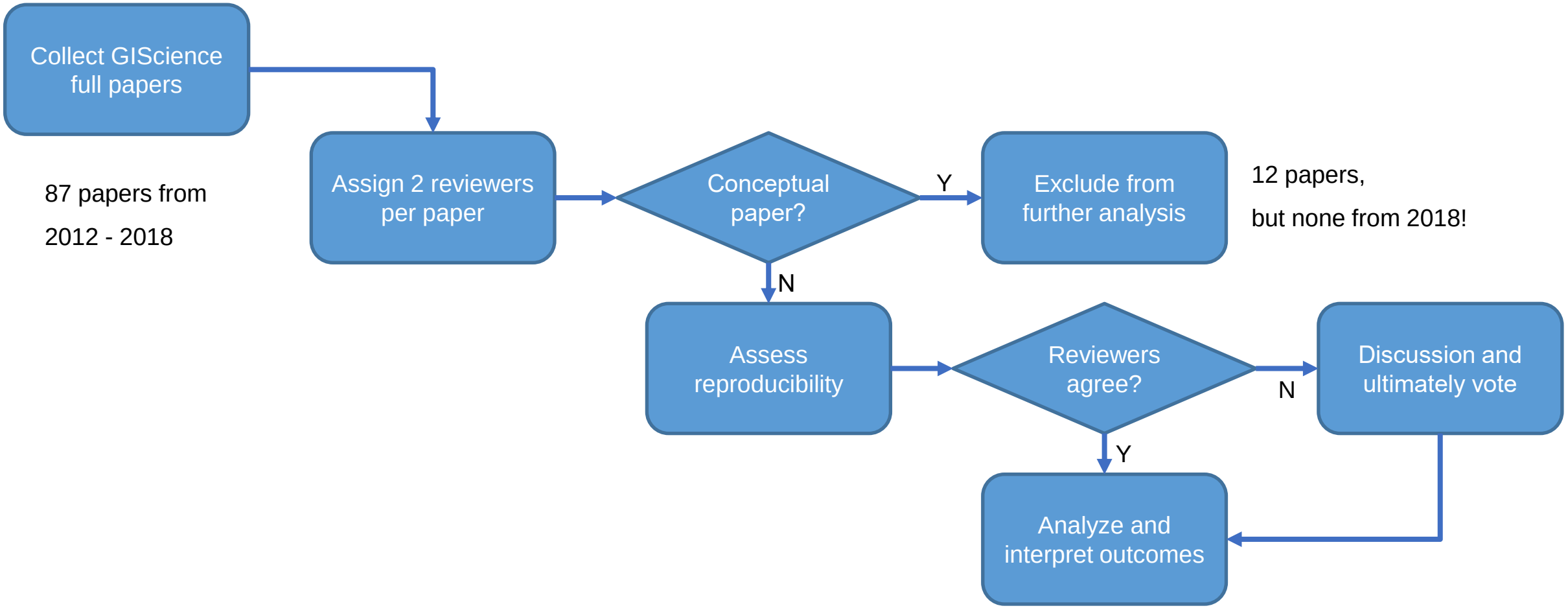
Reproducible Research and / at GIScience

(GIScience: Bi-annual conference series with global target audience)

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?

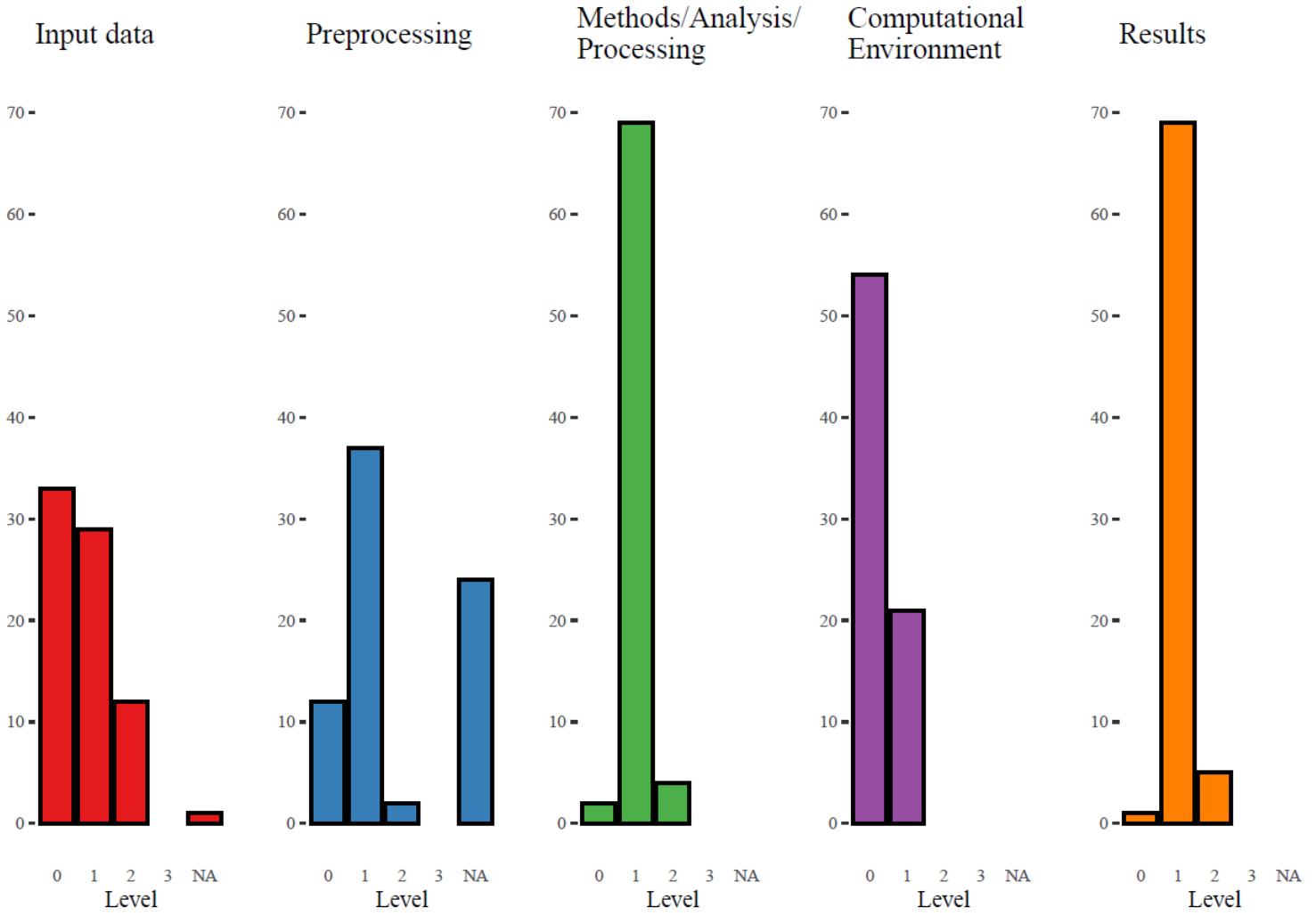


Figure 1 Barplots of reproducibility assessment results; levels range from 0 (leftmost bar) to 'not applicable' (rightmost bar).

Level – Description

0 – unavailable

1 – documented

2 – available

3 – available and open

Any patterns visible?

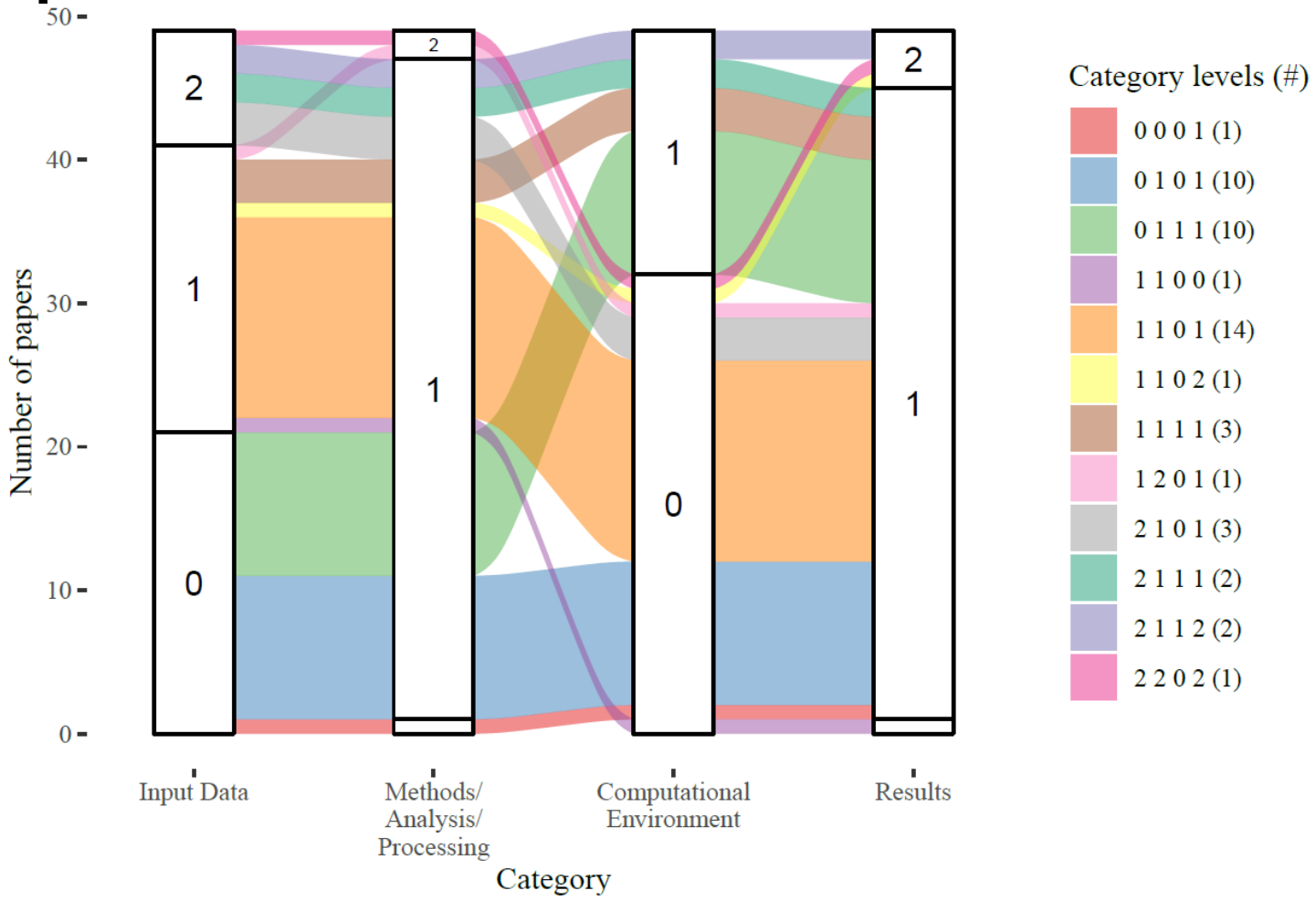
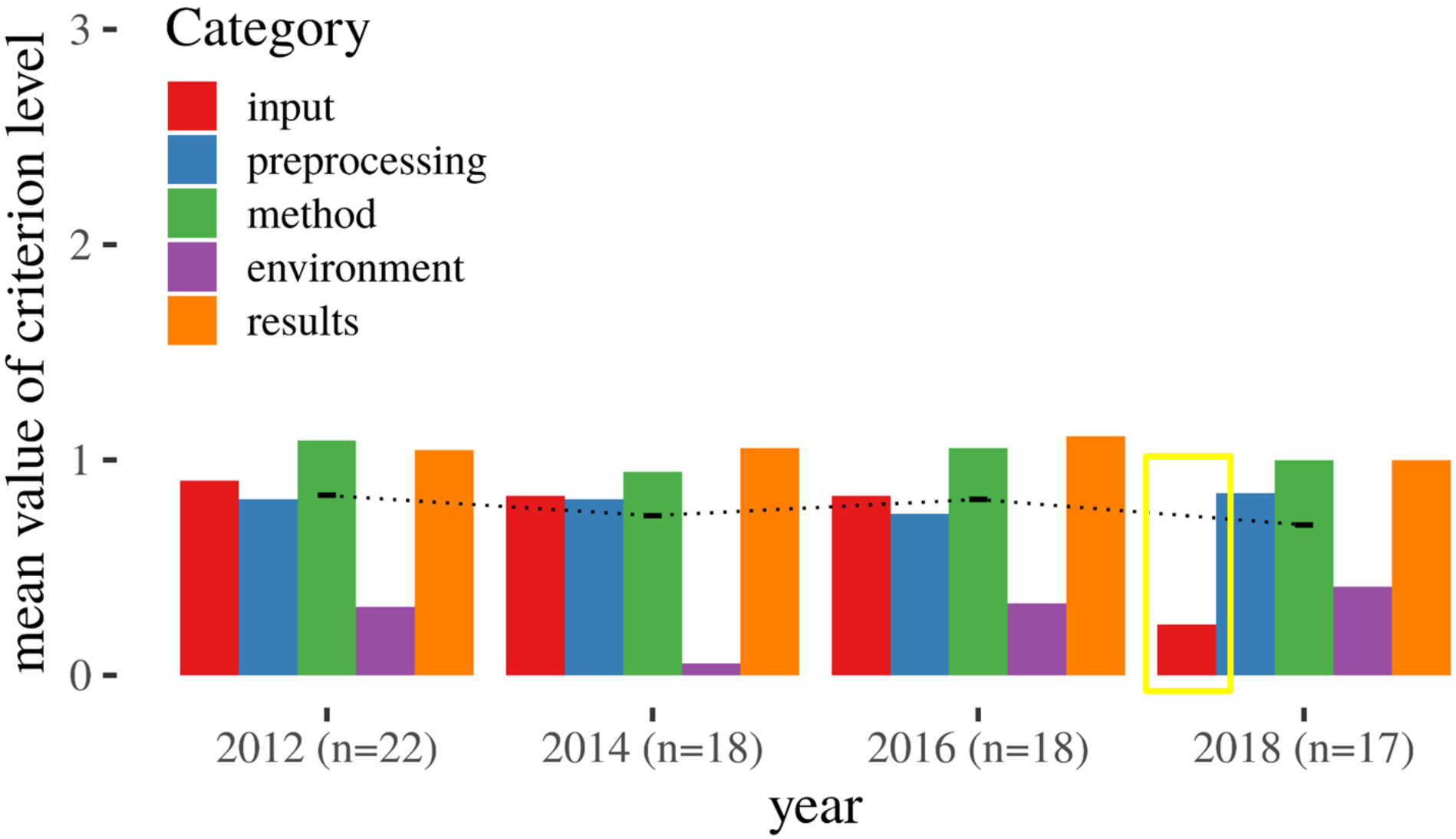


Figure 2 Alluvial diagram of common groups of papers throughout 4 of 5 categories including only papers without any "not applicable" *(Level NA)* value; category *Preprocessing* was dropped because difficulty to clearly assess it lead to many "not applicable" values.

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

What could be options for ~~GIScience~~ conferences?

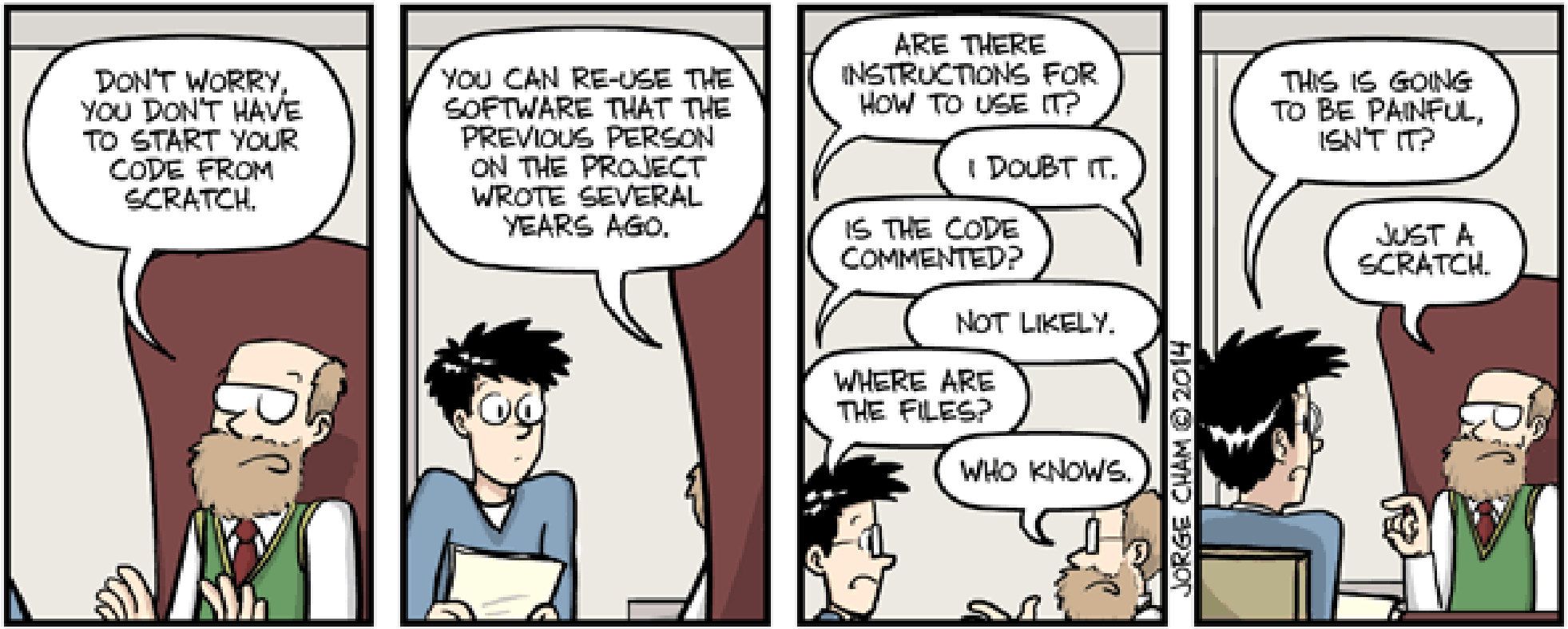
Keeping in mind:

- Reproducibility is not all-or-nothing game
- Culture change can be supported and encouraged, but not forced
- ***Don't exclude studies requiring proprietary software or input data that cannot be shared (privacy!) - but make sure they do their best to be as reproducible as possible***

- Technology seems less of an issue than cultural / community practices
- Reproducibility committee, badges, and joint working group seem difficult to set up and maintain without institutional support
- look at AGILE reproducibility guidelines and adapt and adopt
- make reproducibility a major criterion for review: if authors haven't done everything they are expected (define clear expectations!), then reject the manuscript

Teaching the new researchers – Reproducibility in the classroom

A Senior University Teaching Qualification project



http://phdcomics.com/comics.php?f=1689

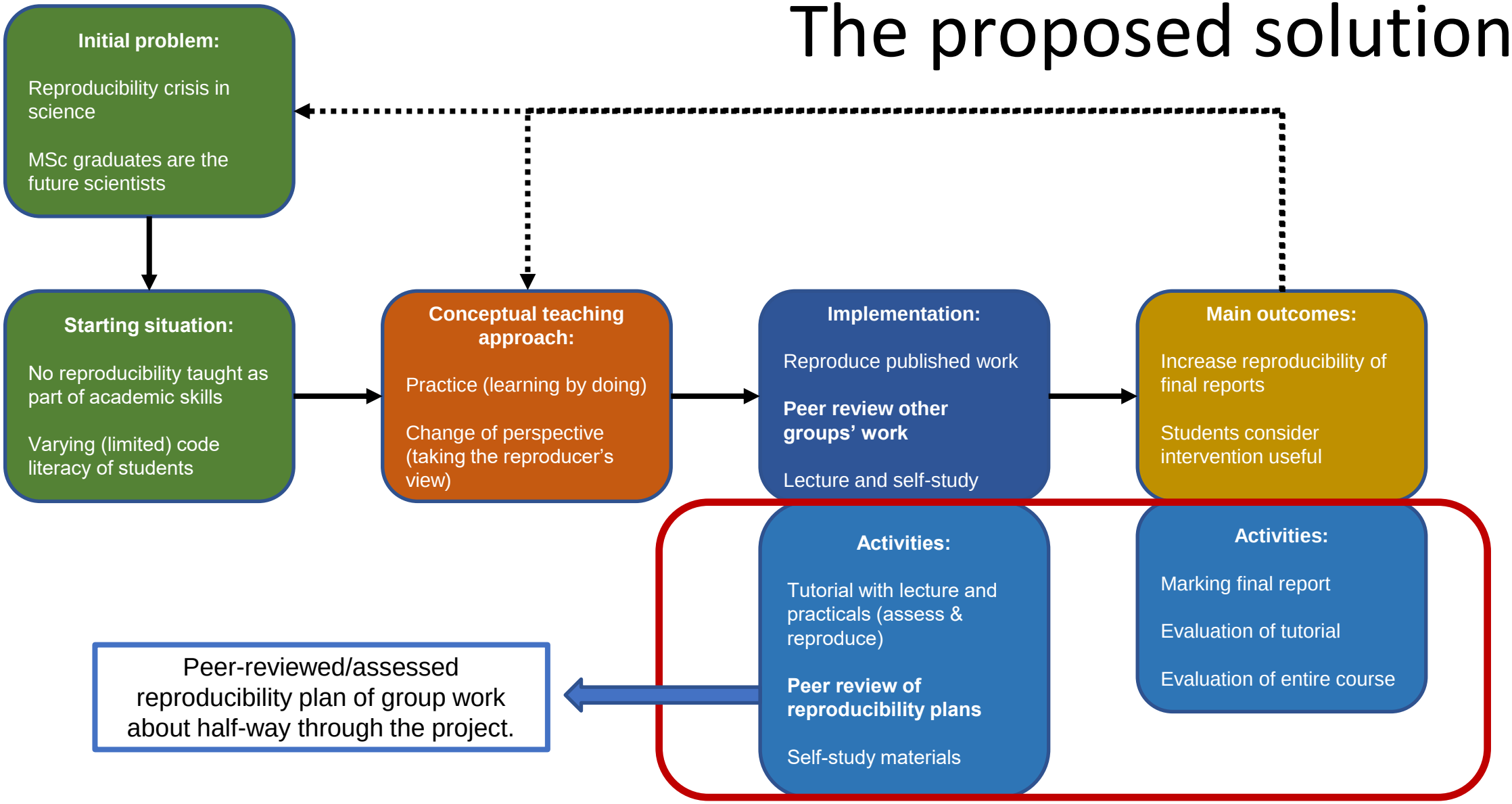
The challenge

- 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

The context

- New MSc program “Spatial Engineering”:
  - Project-based learning
  - Elements of challenge-based learning
  - Wicked problems
  - Process is more important than outcomes
- Senior University Teaching Qualification
  - Promotes investigation of novel / different teaching approaches
  - 18 months to complete
  - Results in course materials and a report

The proposed solution



The results

Evaluation	2019 (pre-intervention)		2020 (intervention)	
	Mean scores (n=4)	teacher assessment	Mean scores (n=5)	teacher assessment
Data	0.75	Links to important data mostly provided, but far from complete	1	Most data is available through links, more information on how data was generated
Methods	0.75	very little concrete information on computational environment, and no code	1	Analysis details often added in an appendix of the assignment reports
Results	0.75	Not clear how specific figures or tables were created	1.4	All results are fully described and linked with analysis steps
Share of reports with at least basic reproducibility	50%	Reproducibility not recognized as an important aspect, although one student group briefly assessed reproducibility of their work.	80%	All but one group submitted a reproducibility plan, and all groups except one reached at “Documented” in all criteria

The students' view

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
… the peer-reviewed reproducibility plan?	**0**	**2**	**7**	**4**
Summary	1	9	27	16

The follow-up

- Intervention ran again with good results
- Will continue and expand in other MSc programs, too
- Effect on final thesis difficult to measure, survey did not work well due to pandemic

WHAT CAN YOU DO TODAY?

- Descriptive and consistent
  - File names
  - Variable names
- Document for future you
- Plain text + version control systems (e.g., git)
- Free and open-source software and formats
- Follow FAIR principles
  - https://www.nature.com/articles/sdata201618
  - https://www.force11.org/fairprinciples
  - https://www.go-fair.org

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.

WHAT CAN YOU DO TODAY?

- Executable digital notebook (e.g. Jupyter, compare https://en.wikipedia.org/wiki/Open-notebook_science)

- 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".

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 that the way notebooks are being used leads to unexpected behavior, encourage poor coding practices, and that their results can be hard to reproduce. To understand good and bad practices used in the development of real notebooks, we studied 1.4 million notebooks from GitHub. We present a detailed analysis of their characteristics that impact reproducibility. We also propose a set of best practices that can improve the rate of reproducibility and discuss open challenges that require further research and development.**

***Index Terms*—jupyter notebook, github, reproducibility**

I. Introduction

Literate programming is a paradigm that seeks to help in the communication of programs [1] by interleaving formatted natural language text, executable code snippets, and computation results. Code snippets generate the computation results and natural language text explains both the code and the results.

Jupyter Notebook is the most widely-used system for interactive literate programming [2]. It was designed to make data analysis easier to document, share, and reproduce. The system was released in 2013, and today there are over 1 million notebooks in GitHub [3]. Jupyter Notebook originated from IPython [4] and, in addition to Python, it supports a variety of programming languages, such as Julia, R, Javascript, and C. Notebooks interleave not only code and text, but also different kinds of rich media, including image, video, and even interactive widgets combining HTML and JavaScript.

Kluyver et al. [5] advocate the usage of notebooks for publishing reproducible research, due to their ability to combine reporting text with the executable research code. However, the format has been increasingly criticized for encouraging bad habits that lead to unexpected behavior and are not conducive to reproducibility [6]–[8]. Among the main criticisms are hidden states, unexpected execution order with fragmented code, and bad practices in naming, versioning, testing, and modularizing code. Also, the notebook format does not encode 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].

Existing work attempted to understand how notebooks are used [3], [13], [14]. They analyzed different aspects of notebooks, including use cases [13], narrative [3], [13], and structure [3], [14]. However, they did not attempt to run the notebooks and check characteristics related to reproducibility.

In this paper, we present a study that aims to provide insights into the reproducibility aspects of real notebooks. To better understand the different characteristics that impact reproducibility, using the aforementioned criticisms as a guide, we define metrics to analyze the extent of adoption of both good and bad practices. To compute these metrics, we created a corpus consisting of 1,159,166 unique notebooks collected from 264,023 GitHub repositories and extracted information about the structure of the notebooks. Besides, to assess the reproducibility rate, we attempted to execute the notebooks. As we discuss in Section IV, out of 863,878 attempted executions of valid notebooks (i.e., notebooks with defined Python version and execution order), only 24.11% executed without errors and only 4.03% produced the same results. Based on our findings, we propose a set of best practices for the development of Jupyter Notebooks.

This paper is organized as follows. Section II provides some background about literate programming and Jupyter Notebooks. Section III describes the method we followed in this study and our notebook corpus. We present the analysis results in Section IV. In Section V, we propose a set of best practices for the development of Jupyter Notebooks. We 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

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, I know how valuable and difficult it is.
- Remember: Reproducibility is a spectrum. Let’s try to make qualitative research as reproducible as possible!
- But how? -> Anyone volunteering to find out?

Thanks a lot for your attention!