Open, reproducible, and replicable (PhD) research

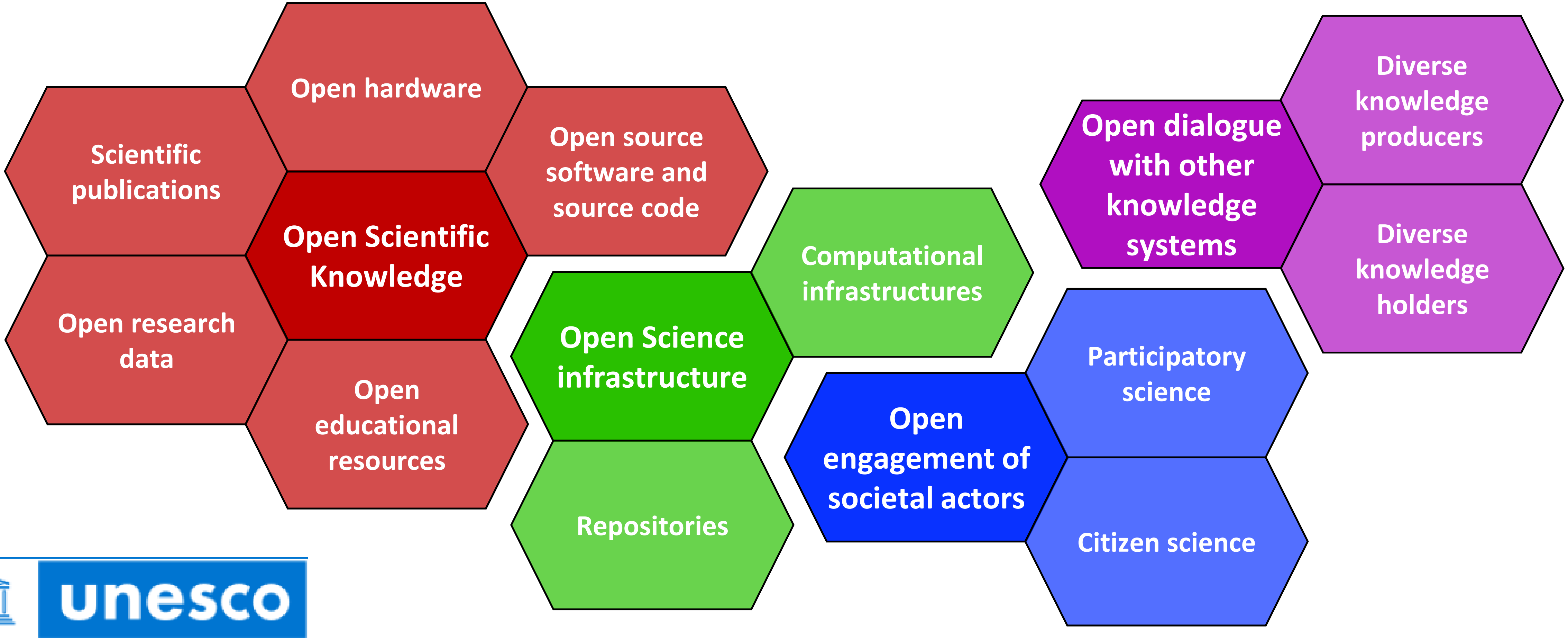
Based on AGILE PhD school 2022

Frank Ostermann, ITC, University of Twente

What does Science mean?

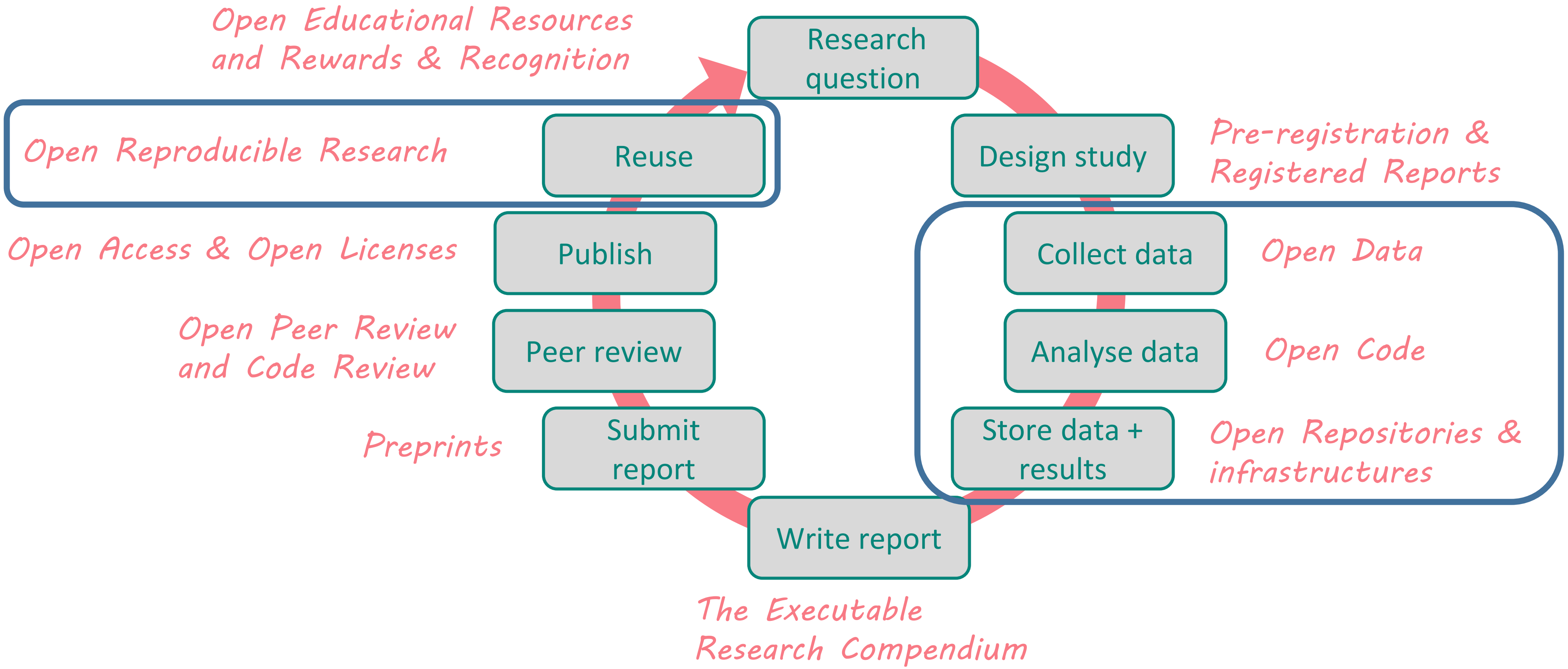
- Discover laws, axioms, rules, etc. and describe under which conditions they apply
- Conduct case studies to prove a general principle or theory
- Transfer/publish results to prove validity, veracity, trust in findings
  - Scientists confirm the validity of a new finding or discovery by **repeating** the research.
  - Observed inconsistency may be an important precursor to new discovery while others fear it may be a symptom of a lack of rigor in science

What does Open Science mean?


Open Scientific Knowledge
Scientific publications
Open hardware
Open source software and source code
Open research data
Open educational resources
Open Science infrastructure
Computational infrastructures
Repositories
Open engagement of societal actors
Participatory science
Citizen science
Open dialogue with other knowledge systems
Diverse knowledge producers
Diverse knowledge holders

Further reading: UNESCO Recommendation on Open Science

What does Open Science mean?



Research question

Design study

Pre-registration & Registered Reports

Collect data

Open Data

Analyse data

Open Code

Store data + results

Open Repositories & infrastructures

Write report

The Executable Research Compendium

Submit report

Preprints

Peer review

Open Peer Review and Code Review

Publish

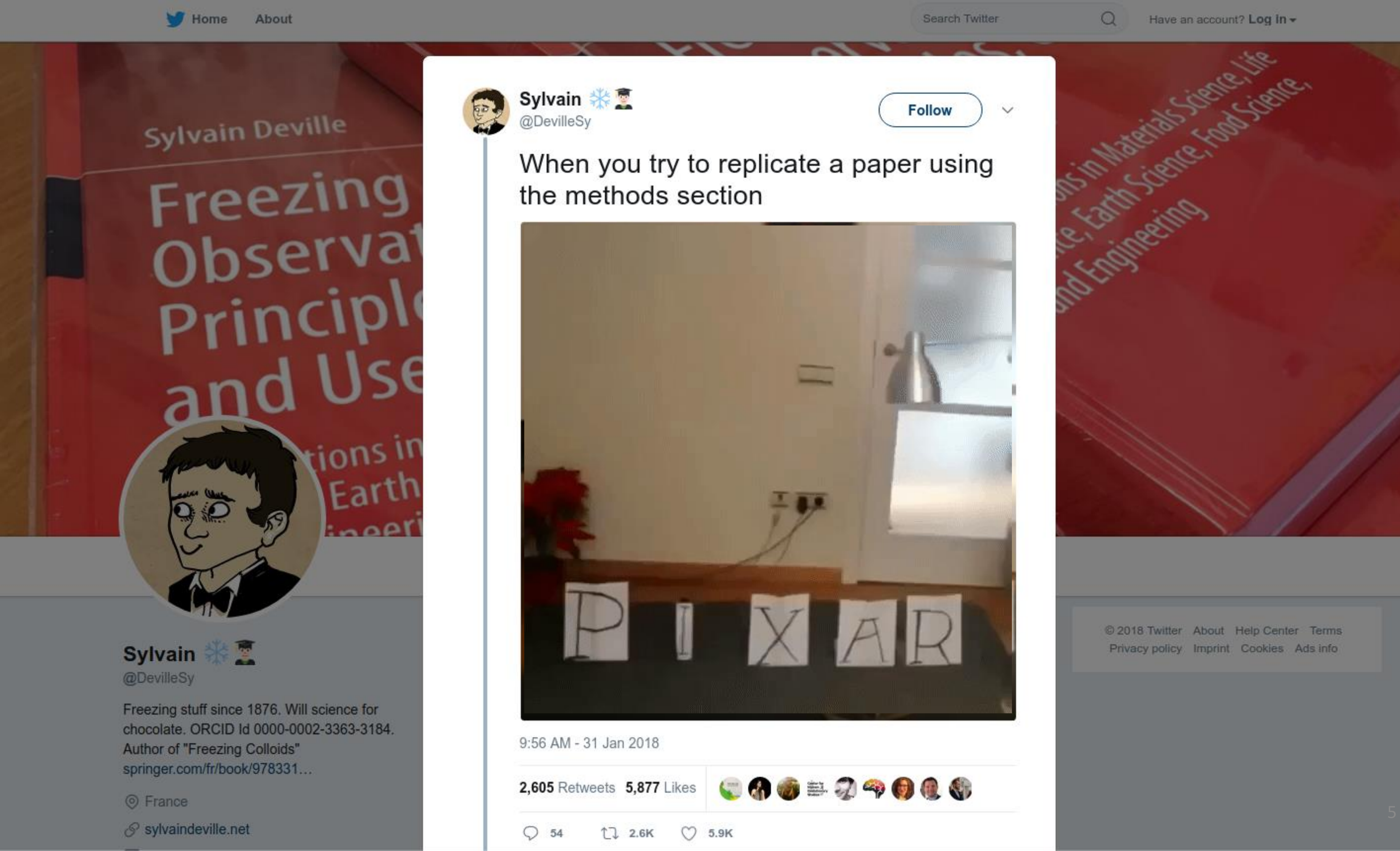
Open Access & Open Licenses

Reuse

Open Reproducible Research

Open Educational Resources and Rewards & Recognition

Original slide by Dr. Markus Konkol; modified



Sylvain ❄️
@DevilleSy

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

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…

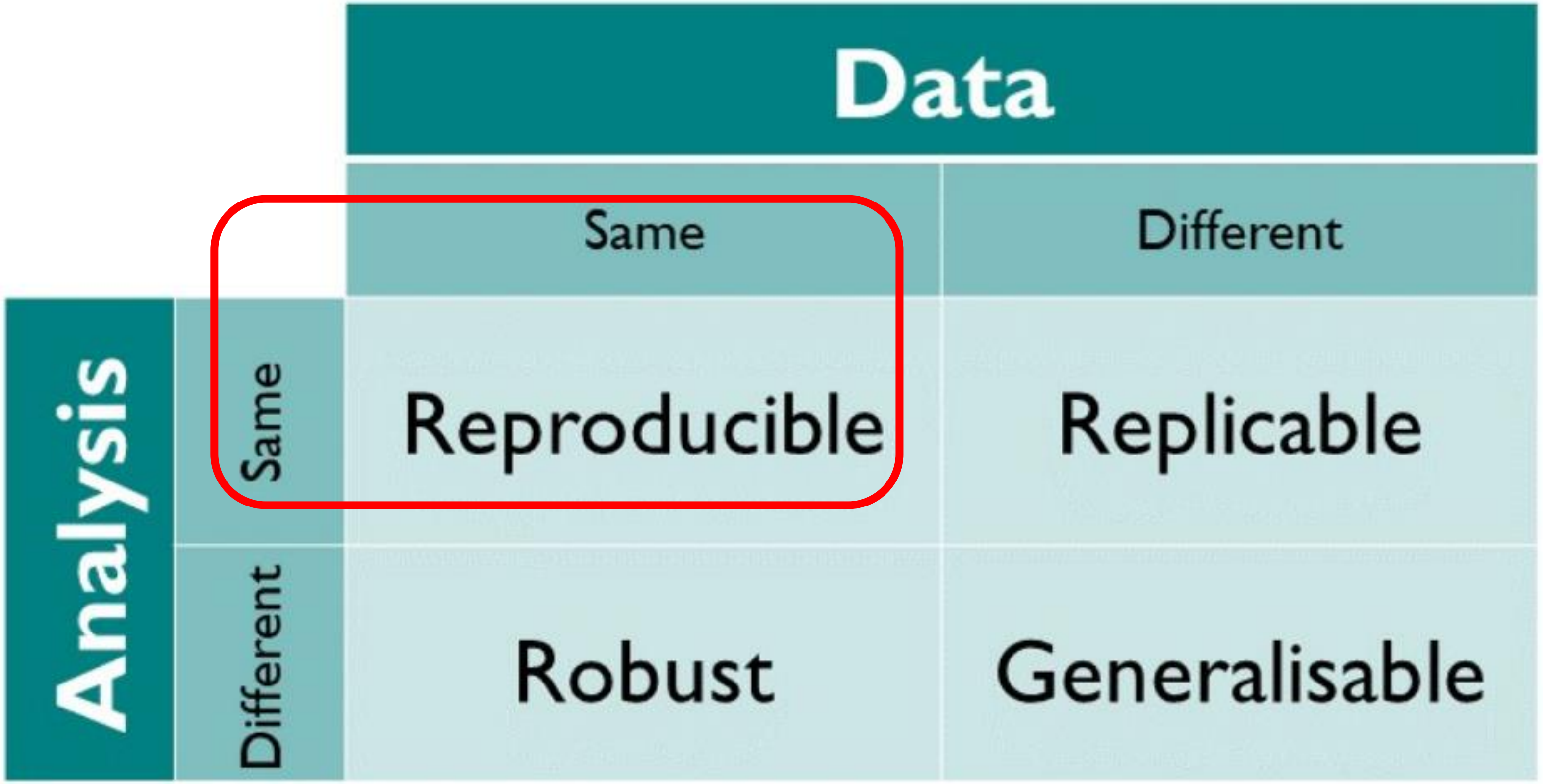
France

sylvaindeville.net

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

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

Reproducible Research



		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

Further reading: [The Turing Way]()

Original slide by Dr. Markus Konkol; modified

What are reproducibility and replicability?

Reproducibility is whether you can get the **same results** using

- the **same analysis** (applying the same methods and code, libraries, programs, etc.) and
- **same original data**

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

What are reproducibility and replicability?

Replicability is whether you can get **similar results** changing

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

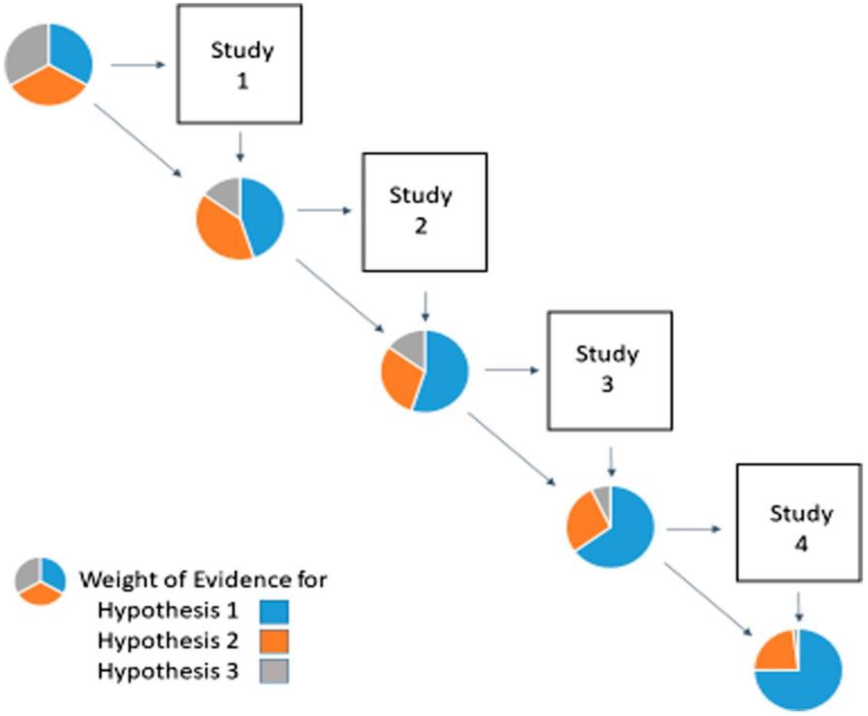
If outcomes are similar, original hypothesis is supported

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

Open != reproducible

“Openness and Open Science (data sharing, code sharing, open access, etc.) are enablers of reproducibility, but do not necessarily guarantee it” https://zenodo.org/record/5521077)

By default

- Open != good (of high academic quality)
- Reproducible != Good (of high academic quality)
- Open != reproducible

What does open and reproducible science mean for PhD (MSc) research?

Start early….

.. And don’t ask for permission to document your analysis (or data, if possible), just make this part of your manuscript/thesis

Reproducible analysis

- R or Python script is best as any one can reproduce your analysis
- If you had to re-do your analysis changing one variable, could you do this quickly?

Make data open

- Complex if privacy considerations applied
- If you cannot, synthetic/simulated datasets are an option

Posting preprints and postprints

You don’t need to adopt all these practices at once

- It’s not all or nothing

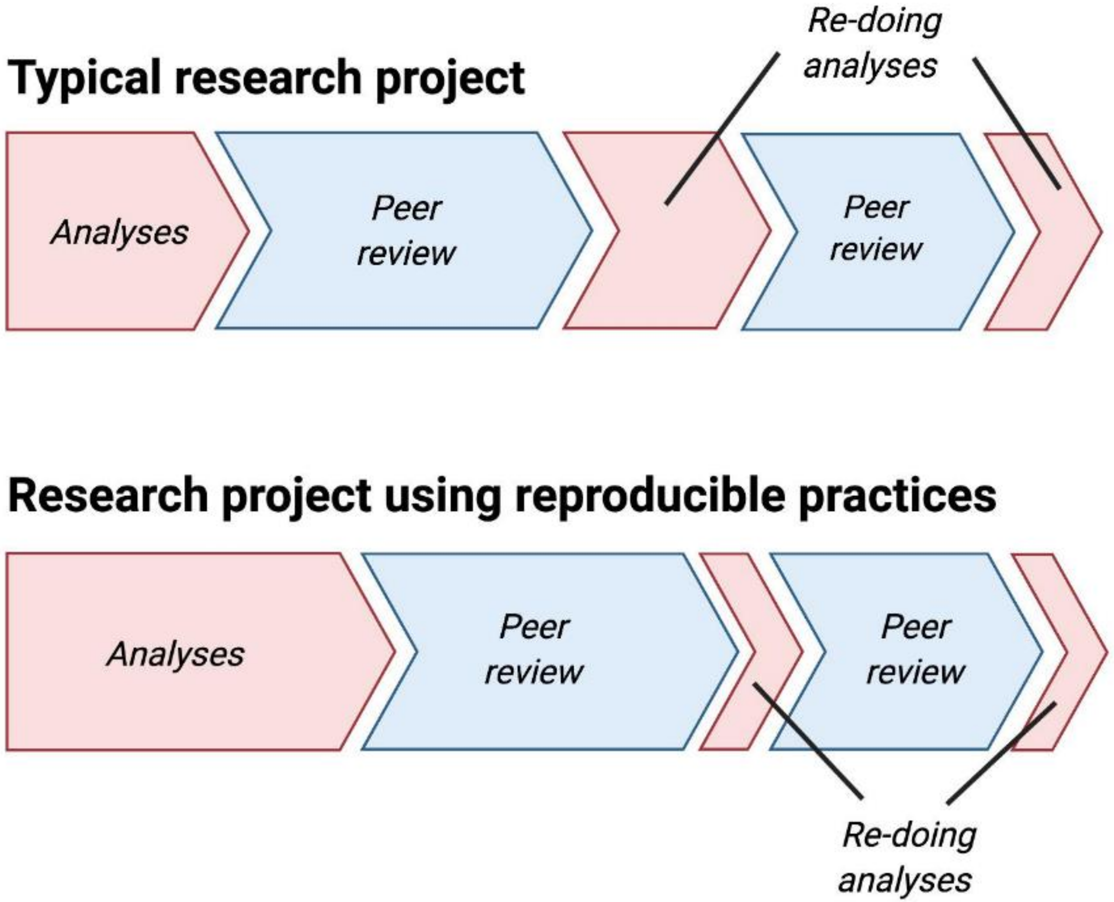
- Start applying practices paper-by-paper. It becomes easier every paper

- Even if no one checks your data and scripts (so far), the very practice of making data and code available builds your credibility (and your open science/data science skills!)

You don’t spend more time doing reproducible science

- You just **realocate**

where you spend it



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

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!

WHAT ARE THE MAIN REPRODUCIBILITY CHALLENGES FOR USING CROWDSOURCED OR VOLUNTEERED GI?

REPRODUCIBILITY

1. Platform (API) Black boxes:

You can’t guarantee that others will retrieve the same data

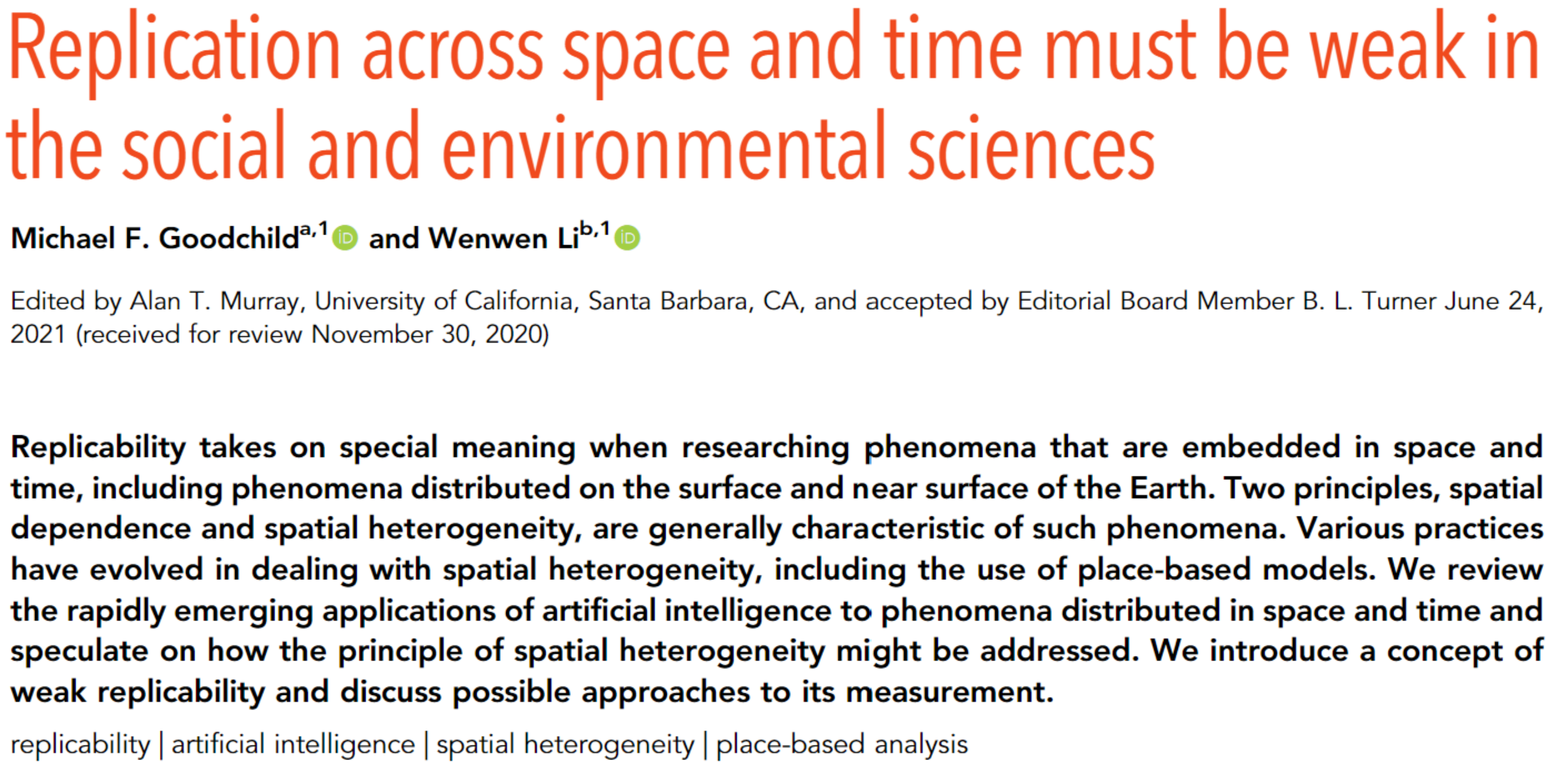
2. Volatility of content and access:

You can’t guarantee that the content will remain the same, nor that others will continue to be able to access it (licenses, ToS)

3. Variance in human behavior (inter- and intra-rater agreement):

You can’t guarantee that volunteer data is consistent, even from one participant

Qualitative (GIScience) research



Replication across space and time must be weak in the social and environmental sciences

Michael F. Goodchild[a,1] **and Wenwen Li**[b,1]

Edited by Alan T. Murray, University of California, Santa Barbara, CA, and accepted by Editorial Board Member B. L. Turner June 24, 2021 (received for review November 30, 2020)

Replicability takes on special meaning when researching phenomena that are embedded in space and time, including phenomena distributed on the surface and near surface of the Earth. Two principles, spatial dependence and spatial heterogeneity, are generally characteristic of such phenomena. Various practices have evolved in dealing with spatial heterogeneity, including the use of place-based models. We review the rapidly emerging applications of artificial intelligence to phenomena distributed in space and time and speculate on how the principle of spatial heterogeneity might be addressed. We introduce a concept of weak replicability and discuss possible approaches to its measurement.

replicability | artificial intelligence | spatial heterogeneity | place-based analysis

https://www.pnas.org/doi/10.1073/pnas.2015759118

Qualitative (GIScience) research – 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