Contact
Martin Schweinberger
The University of Queensland, Australia
E: m.schweinberger@uq.edu.au
The Arctic University of Norway, Tromsø
E: martin.schweinberger@uit.no
H: www.martinschweinberger.de

Best Practices in Corpus Linguistics

What lessons should we take from the Replication Crisis and how can we guarantee high quality in our research?

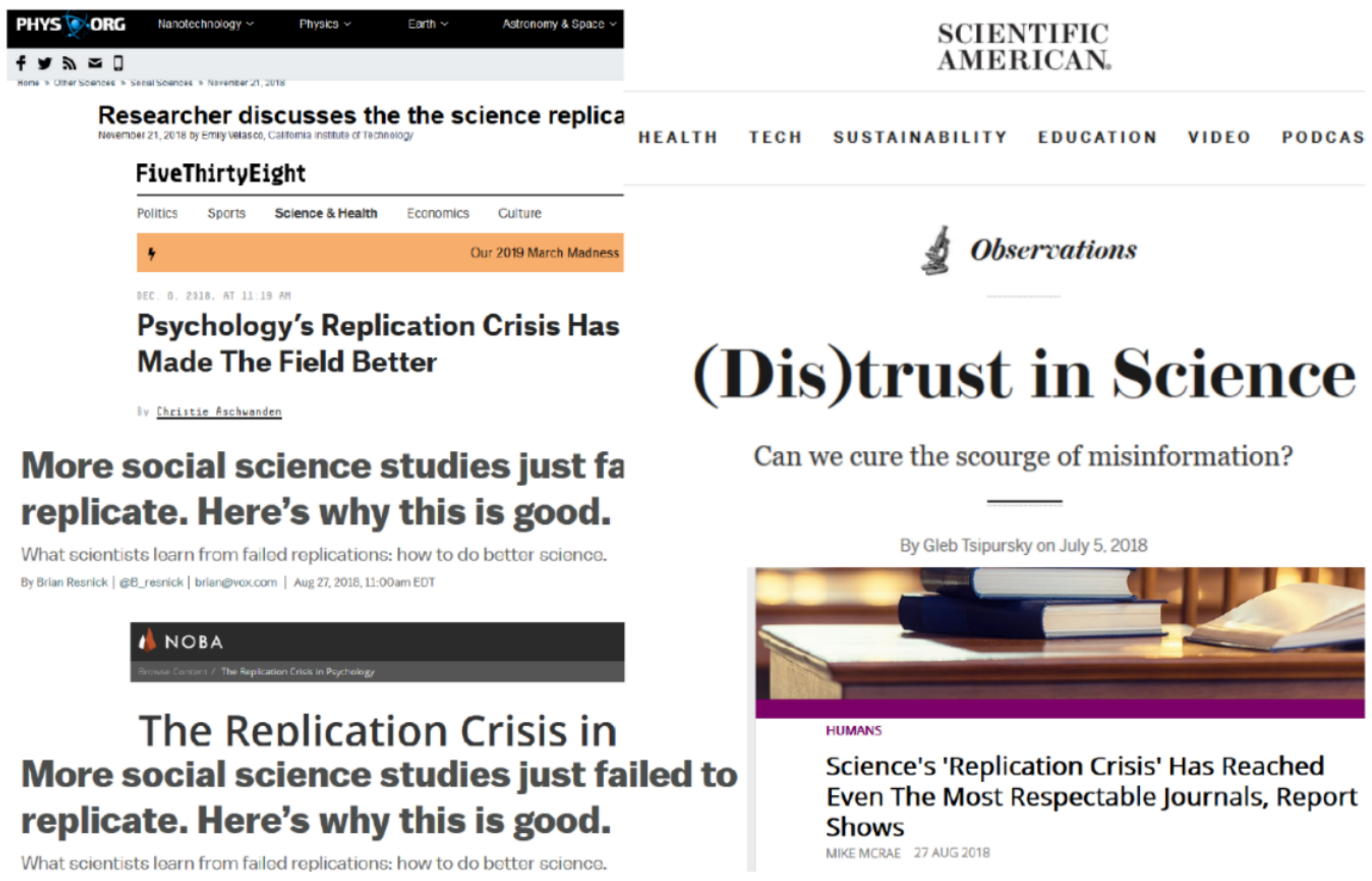
Aims, definition, and the current state of affairs

This poster aims to
- Raise awareness for ***Best Practices*** (BP) in Corpus Linguistics
- Discuss issues related to Reproducibility and Replicability
- Propose improvements to current research practices
- Offer solutions on how best practices can be implemented

A BP is a method or technique that is superior to alternatives because it produces results that are more reliable, transparent, replicable, and in compliance with legal or ethical requirements.

BP have come into focus as a result of the **Replication Crisis** (RC) which is a ongoing methodological crisis primarily affecting parts of the social and life sciences beginning in the early 2010s.

Nature 2016 poll of 1,500 scientists:
- 70% failed to reproduce at least one other scientist's experiment
- 50% failed to reproduce one of their own experiments (Fanelli 2009)



PHYS ORG
Researcher discusses the the science replica
November 21, 2018 by Emily Velasco, California Institute of Technology

FiveThirtyEight
DEC. 8, 2018, AT 11:19 AM
Psychology's Replication Crisis Has Made The Field Better
By Christie Aschwanden

More social science studies just fa
replicate. Here's why this is good.
What scientists learn from failed replications: how to do better science.
By Brian Resnick | @B_resnick | brian@vox.com | Aug 27, 2018, 11:00am EDT

NOBA
The Replication Crisis in

More social science studies just failed to replicate. Here's why this is good.
What scientists learn from failed replications: how to do better science.

SCIENTIFIC AMERICAN
Observations
(Dis)trust in Science
Can we cure the scourge of misinformation?
By Gleb Tsipursky on July 5, 2018

HUMANS
Science's 'Replication Crisis' Has Reached Even The Most Respectable Journals, Report Shows
MIKE MCRAE 27 AUG 2018

Examples for media outlets reporting on the Replication Crisis.

As a consequence of the RC, there is growing awareness. . .
- of a problem: currently most research is difficult to replicate/reproduce!
- that reproducibility is an essential part of the scientific method
- that the inability to replicate has potentially grave consequences as significant theories are grounded on unreproducible work
- that there is substantial loss of trust in science, its results, and its proponents.

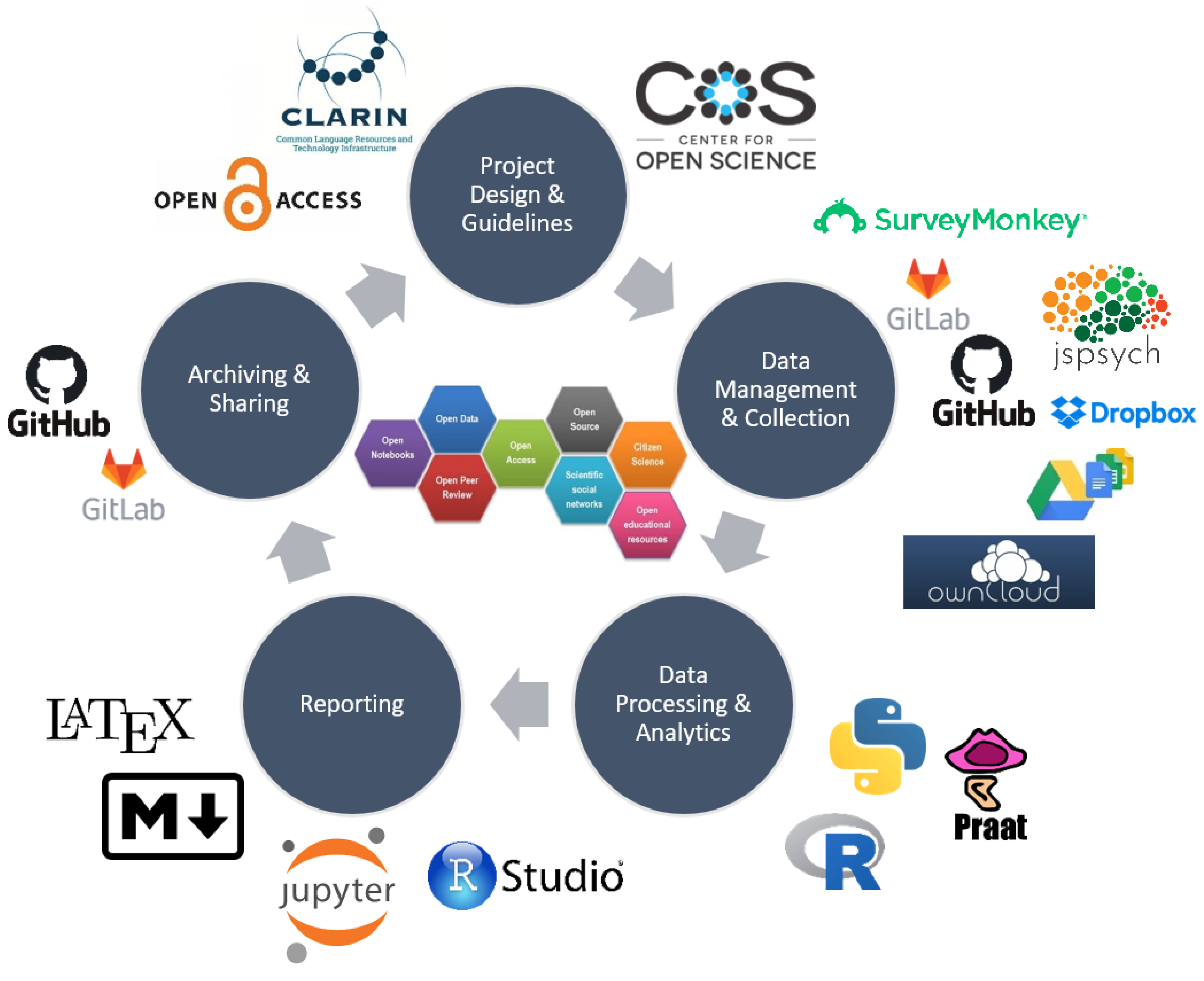
Best Practices and Replication in Corpus Linguistics

As a community, we endorse *blind peer-review*, we are *open to sharing* (if we are asked), and we have *begun with a discussion around BP* and replication (Berez-Kroeker et al. 2018; Ruhi et al. 2014).

However, we could be better because *our analyses are not reproducible*, we have an *over-reliance on tools*, and *reproduction is discouraged* ((i) journals are not interested in publishing the same analysis twice; (ii) researchers fear repercussions if they criticize the research of others (face-threatening).

While replicability has improved with the rise of natural language corpora, **we just do not know how bad our research is** (mistakes in using statistical methods or data processing, outright forgery, data manipulation, p-hacking, etc.) because . . .
1. researchers do not (or only rarely) reproduce and replicate
2. researchers do not know about BP or what they are
3. researchers do not know how to make their research comply with BP
4. lack of training in BP and how to make research reproducible

Also, **BP** make research not only more transparent and reproducible but they also **help in optimizing work flows, thus making research more efficient** and help saving time because (part of) scripts can be re-used and because your **files will be much tidier!**

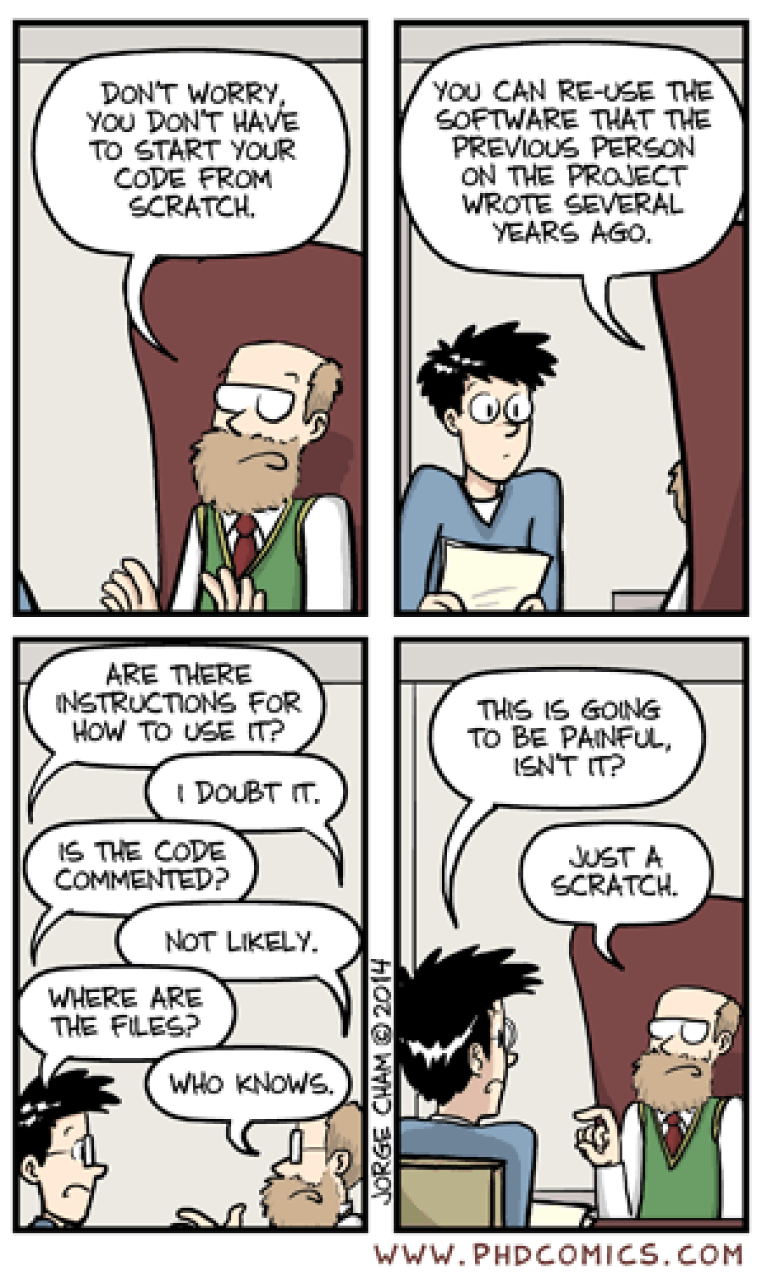

CLARIN
Common Language Resources and Technology Infrastructure
COS CENTER FOR OPEN SCIENCE
OPEN ACCESS
Project Design & Guidelines
SurveyMonkey
GitLab
jspsych
Data Management & Collection
GitHub
Dropbox
ownCloud
Archiving & Sharing
GitHub
GitLab
Open Data
Open Notebooks
Open Peer Review
Open Access
Open Source
Scientific social networks
Citizen Science
Open educational resources
LATEX
Reporting
Data Processing & Analytics
jupyter
R Studio
R
Praat

Open science circle with software options to make research more reproducible.

Suggestions to make our research more replicable

For individual researchers and teams

- **FAIR principles**: share and make your data ***F**indable*, ***A**ccessible*, ***I**nteroperable*, ***R**eusable* (FAIR) (Wilkinson et al. 2016)
- **Data a publication**: assign a *Digital Object Identifier* (DOI) to your data, provide a clear example for how your data should be cited, and publish it on an online repository (this way your data is a proper publication)
- **Archiving**: use online repositories (e.g. *GitHub*, *GitLab*, *CouldStore*, *MyDrive*, *Dropbox*) to avoid data loss and various versions of a single document or file
- **Scripts over tools**: use R rather than ready-made software tools because such apps are black-boxes that hinder replication and transparency (due to limited accessibility and/or time-consuming replication)
- **Documentation**: write down what you do and where you store all relevant elements of your project
- **Folder templates**: think about a schematic folder structure and use it for all your projects, e.g. always using subfolders for *data*, *tables*, and *images* for research projects or *slides*, *exercises*, *student materials*, and *assignments* for courses (ideally implement a policy in your team so that all team members use the same folder template)
- **Notebooks and Git**: make your research fully transparent and reproducible by using R or Jupyter Notebooks and sharing entire projects on GitHub or GitLab.


DON'T WORRY, YOU DON'T HAVE TO START YOUR CODE FROM SCRATCH.
YOU CAN RE-USE THE SOFTWARE THAT THE PREVIOUS PERSON ON THE PROJECT WROTE SEVERAL YEARS AGO.
ARE THERE INSTRUCTIONS FOR HOW TO USE IT?
I DOUBT IT.
IS THE CODE COMMENTED?
NOT LIKELY.
WHERE ARE THE FILES?
WHO KNOWS.
THIS IS GOING TO BE PAINFUL, ISN'T IT?
JUST A SCRATCH.
JORGE CHAM © 2014
WWW.PHDCOMICS.COM

For the community

- Endorse *Open Science*:
  Open Data + Open Access + Open Methodology + Open Educational Resources
- Only accept papers that have made data (and scripts) available
- Require data to be cited appropriately (serves as a reward and an incentive to publish corpora)
- Promote replication and support publishing replication studies
- Invest in and support training for staff and students in data management and other options that help make research more transparent (*R*, *Git*, *Markdown*, *wikis*, etc.)
- Continue the discussion and talk to colleagues about *BP*/*Replication*/*Reproducibility*

References

Berez-Kroeker, A. L., L. Gawne, S. S. Kung, B. F. Kelly, T. Heston, G. Holton, P. Pulsifer, D. I. Beaver, S. Chelliah, S. Dubinsky, et al. (2018). Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics 56*(1), 1–18.

Fanelli, D. (2009). How many scientists fabricate and falsify research? a systematic review and meta-analysis of survey data. *PLoS One 4*, e5738.

Ruhi, Ş., M. Haugh, and T. Schmidt (2014). *Best practices for spoken corpora in linguistic research.* Cambridge: Cambridge Scholars Publishing.

Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific data 3*. https://www.nature.com/articles/sdata201618.