From the darkness to the light: reproducibility, replication and transparency in Corpus Linguistics?

Martin Schweinberger

m.schweinberger@uq.edu.au

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

CREATE CHANGE

LADAL

NORGES ARKTISKE UNIVERSITET • UIT •

Logosymbol.

AcqVA Aurora Lab

Background and Motivation

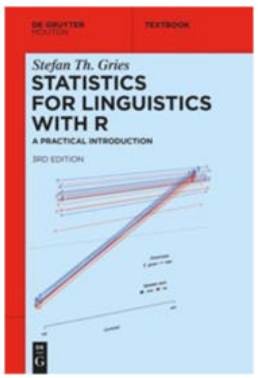
Experiences from consultation (statistics, designs | tools, data management)

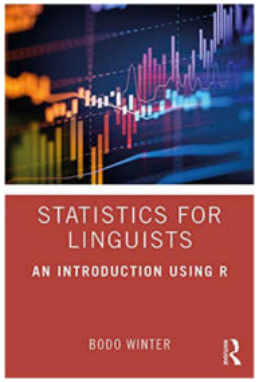
- Everybody's computer is a mess!
  - No | little training (in data management)
- Issues people struggle with
  - Identifying what someone needs is the first step
  - Workflows are often unnecessarily convoluted
  - Untidy data (Organizing data | workflows)
  - Statistics!
  - Automating labour | time intensive tasks

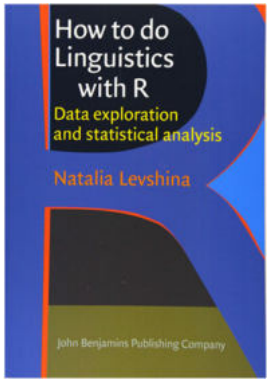
Background and Motivation

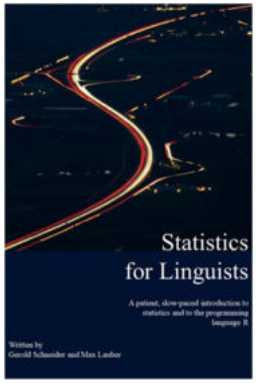
Experiences from consultation (statistics, designs | tools, data management)

- Statistics: myriad of resources
  - Books, Workshops, Bootcamps, Online resources (YouTube | StackOverflow | Quick-R, etc.)
- What about data management | coding and annotation | reproducibility?
  - Resources??? (especially for **qualitative** work!)









Blind Spot: Data Management | Reproducibility | Transparency ?!?

Outline

- Background and Motivation
  - Replication Crisis | Issue | Problem
- Reproducibility, Replication, Transparency
- Options for more transparency (in CL)
  - Practical tips
  - Training infrastructures (LADAL)
- Problems associated with increased transparency (in CL)
- Discussion and Outlook

Background | Motivation

Replication Crisis | Issue | Problem

NOBA

Browse Content / The Replication Crisis in Psychology

The Replication Crisis in Psychology

By Edward Diener and Ro
University of Utah, Univer

AMERICAN PSYCHOLOGICAL ASSOCIATION

MEMBERS TOPICS PUBLICATIONS & DATABASES PSYCHOLOGY HELP CENTER NEWS & EVENTS

Home // Monitor on Psychology // 2015 // 10 // A reproducibility crisis?

A reproducibility crisis?

The headlines were hard to miss: Psychology, they proclaimed, is in crisis.

October 2015, Vol 46, No. 9
Print version: page 39

sciencealert

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.

By Brian Resnick | @B_resnick | brian@vox.com | Aug 27, 2018, 11:00am EDT

SCIENTIFIC AMERICAN

SUSTAINABILITY EDUCATION VIDEO PODCAS

UMANS

Observations

(Dis)trust in Science

Can we cure the scourge of n

By Gleb Tsipursky on July 5

cience's 'Replication Crisis' Has Reached
ven The Most Respectable Journals, Report
hows

IKE MCRAE 27 AUG 2018

FiveThirtyEight

Politics Sports Science & Health Economics Culture

Our 2019 March Madness

DEC. 6, 2018, AT 11:19 AM

Psychology's Replication Crisis Has Made The Field Better

By Christie Aschwanden

PHYS.ORG Nanotechnology Physics Earth Astronomy & Space Technology

Home > Other Sciences > Social Sciences > November 21, 2018

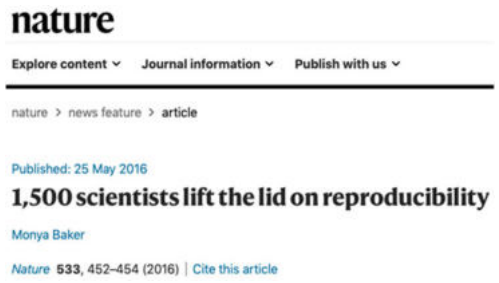
Researcher discusses the the science replication crisis

November 21, 2018 by Emily Velasco, California Institute of Technology

Replication Crisis | Issue | Problem

Controversial ongoing methodological crisis that originated in medicine (Ioannidis 2005) and swiftly expanded to STEM, the social sciences, and psychology when **replications of seminal experiments failed** - calling into question the reliability of widely accepted published research

Reproducibility is a defining feature of science, but the extent to which it characterizes current research is unknown.
(Open Science Collaboration 2015)


nature

Explore content | Journal information | Publish with us

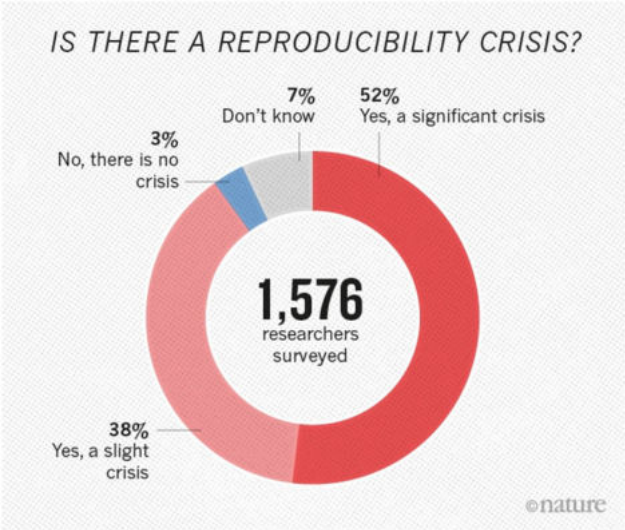
nature > news feature > article

Published: 25 May 2016

1,500 scientists lift the lid on reproducibility

Monya Baker

Nature 533, 452–454 (2016) | Cite this article



"More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments."
(Baker 2016: 452)

Replication Crisis | Issue | Problem

Results and **Effects**

- Public **loss of trust** in science
- Substantive **efforts to improve** transparency and reproducibility (in STEM and “hard” social sciences)
- Examples: increased efforts to support replication, pre-registration, and establishing a culture of sharing & infrastructures for sharing (OSF, GitHub, RNotebooks)

Repercussions of the Replication Crisis in CL

Problem has been identified in (Corpus Linguistics) (recently)

- Workshops
  - ISLE 5 (London, 17-20 July, 2018, L. Sönning & V. Werner “The ‘quantitative crisis’, cumulative science, and English linguistics”
  - ISLE 6 (Joensuu, 2–5 June, 2021): M. Schweinberger & J. Flanagan “Replication and Reproducibility in English Corpus Linguistics”
  - ICAME 42 (Dortmund, 18-21 Aug. 2021): M. Schweinberger, G. Schneider & J. Flanagan “Exploring Powerful Tools to Ensure Robust and Reproducible Results in Corpus Linguistics”

- Journal publications: *Linguistics* 59.5 (Sönning & Werner 2019)

Repercussions of the Replication Crisis in CL

Problem has been identified in (Corpus Linguistics) **BUT focus on data** (data citation, data sharing) **NOT analyses**

Open Access Veröffentlicht von De Gruyter Mouton 6. Dezember 2017

Reproducible research in linguistics: A position statement on data citation and attribution in our field

Andrea L. Berez-Kroeker, Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, David I. Beaver, Shobhana Chelliah, Stanley Dubinsky, Richard P. Meier, Nick Thieberger, Keren Rice und Anthony C. Woodbury

Aus der Zeitschrift Linguistics

see Marsden & Bolibaugh 2021 (to the right), document available through:
https://pure.york.ac.uk/portal/en/publications/reproducibility-and-research-integrity-in-applied-linguistics

This preprint contains the text of a submission of written evidence to the UK Parliament, House of Commons Science and Technology Committee inquiry on reproducibility and research integrity (submitted: 24 September 2021. Viewable on the parliament website here). It is not peer-reviewed.

Reproducibility and research integrity in applied linguistics

Professor Emma Marsden, University of York, emma.marsden@york.ac.uk
Dr Cylcia Bolibaugh, University of York, cylcia.bolibaugh@york.ac.uk

We work in the area of applied linguistics, with a focus on the learning of languages (second, foreign, additional languages after the first language). This is a multidisciplinary field, sitting at the intersection of social sciences (education), arts & humanities (linguistics, languages) and learning sciences (psychology, including neuroscience).

We are writing in our capacity as Director (Emma Marsden) and Co-Director (Cylcia Bolibaugh) of two open research and impact initiatives: IRIS (Instruments and materials for Research Into Second languages) and OASIS (Open Accessible Summaries In Language Studies).

Availability of data and code underpinning published findings

Sharing of data and code underpins computational reproducibility, and is necessary for the verification of individual studies, as well as for the carrying out of meta-analyses. Failure to share data results in a cumulative loss of research value as findings cannot be incorporated into research syntheses and meta-analyses.

Reproducibility | Replication | Transparency

Replication Crisis | Issue | Problem

Reproducibility

- To reproduce a study means doing the things to the same data to get the exact same results.

Replication

- Replicating a study means doing the same | (similar) things to similar data

Robustness (**Generalizability**, National Science Foundation 2018)

- Robustness | Generalizability refers to results being consistent | stable across replications

Replication Crisis | Issue | Problem

Practical vs theoretical reproducibility

- **Practical** reproducibility means that **reproducibility is made easy** for researchers given existing constraints (time, skills, technology, copyright, etc.)
- **Theoretical** or formal reproducibility means that **reproduction is possible** in principle but hindered by real-world restrictions (data only accessible in a specific lab | study based on black box tools or is accompanied by spreadsheets not code)

Replication vs Transparency

But do we really want Reproducibility?

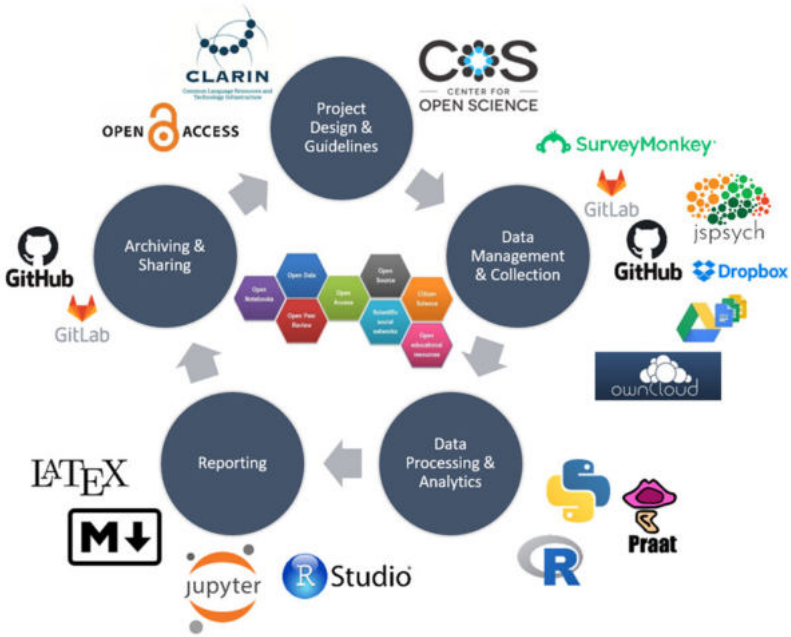
- Difference between reproducibility across different fields
  - *software development*: focus on **technical** aspects
  - *linguistics*: **conceptual** reproducibility
- As reviewers and researchers, we want to understand and be able to check annotation (inspect how the researcher has coded individual instances of language use)
- **Choices and decisions should be transparent** (using a log/notebook) rather than technical reproducibility

Problem | Issue

Even if researchers want to be more transparent, there are limited broad-range resources for HASS researchers that show how to document their research on textual data in a transparent manner!

Options for more transparency (in CL)

How to improve reproducibility | replicability | robustness

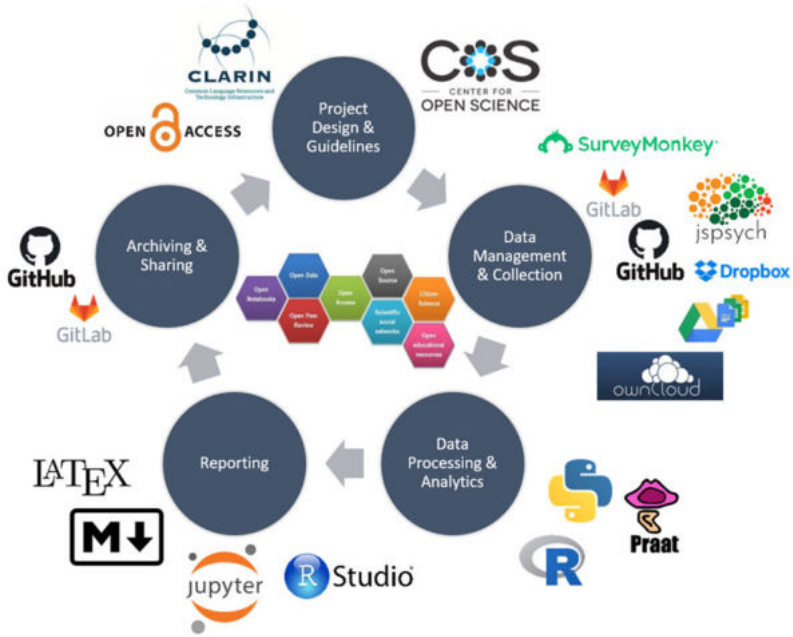


Data management

- **File naming**: consistent and meaningful
- **Folder templates**: Use templates across teams | labs
- **3-2-1 rule**: 3 copies of data on 2 media one of which should be the cloud
- **Documentation**:
  - document where to find what
  - helpful for on-boarding people
  - useful when sharing projects
  - allows to recover what has been done and helps avoiding data loss (bus factor): how many people can be run over by a bus without the project coming to a halt?)

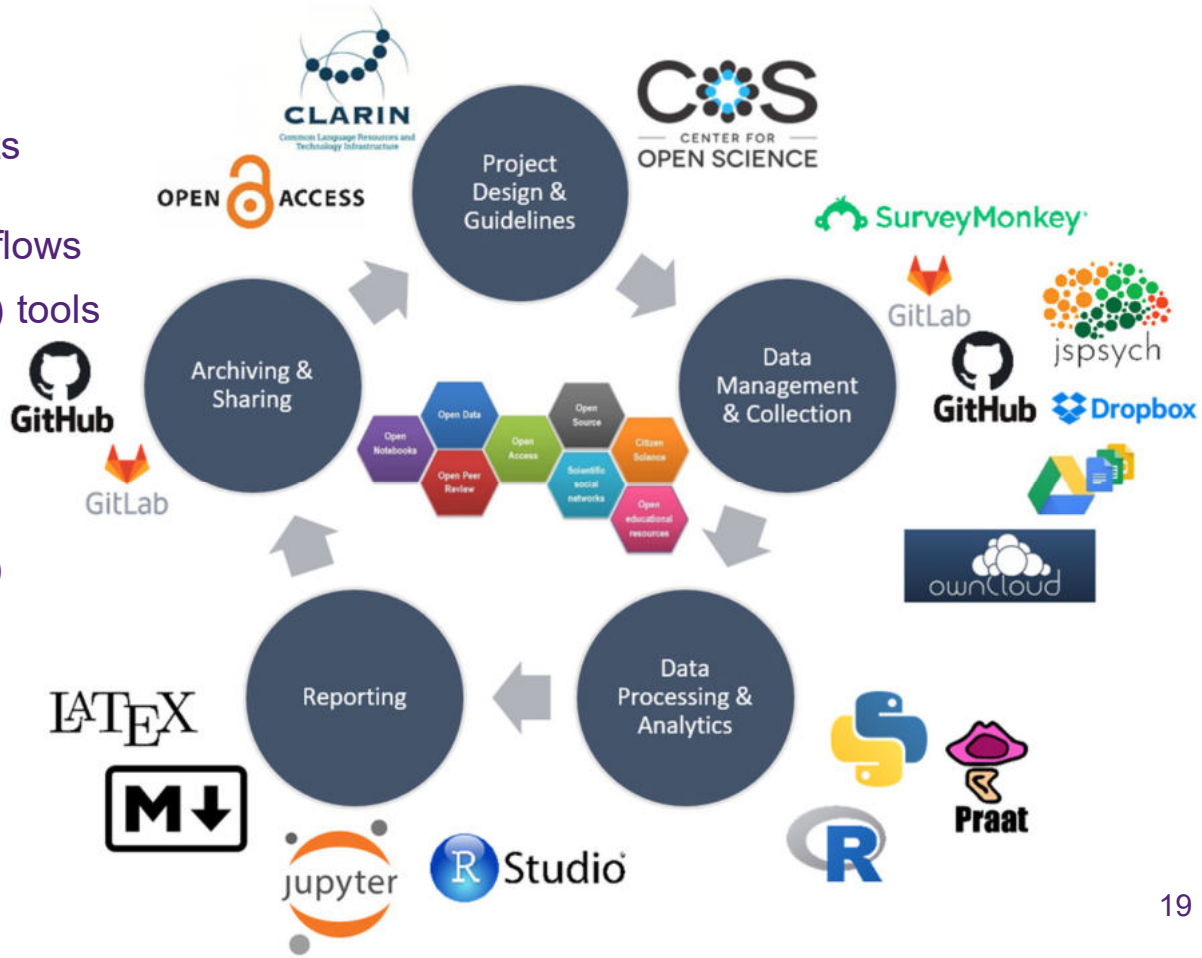
How to improve reproducibility | replicability | robustness

- **FAIR data**
  - Findable, Accessible, Interoperable, Reusable
  - Sharing data (OSF, GitHub, etc.)
- **Practice shift**
  - Pre-registration
  - Upskilling for MA students
  - Stronger focus on replication studies
  - Submitting notebooks | code & data alongside papers
  - Acknowledge data sets as research outputs



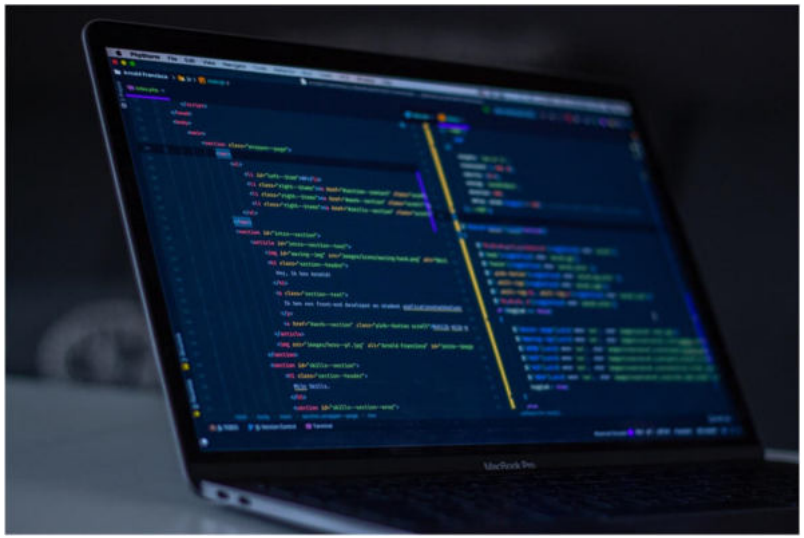
How to improve reproducibility | replicability | robustness

- **Transparency | Reproducibility**
  - Version control and contained environments (Git, renv, conda, Binder, Docker)
  - Notebooks (Rmd, Jupyter): recording workflows
  - Commented scripts rather than (blackboxy) tools
  - Sharing work: OSF, GitHub, GitLab
- **Tools**
  - Using fewer tools
  - RStudio (Rproj, Rmd, renv, Git-integration)
  - Aarnet's SWAN (?)
- **Infrastructures | Training**
  - LADAL
  - ATAP
  - Sydney Corpus Lab



Challenges for infrastructures

- Audiences with very different **levels of expertise**
- **Audiences** with vastly different interests, expectations, and needs
- Training is required at **different levels of specificity** (general introductions vs highly specific methods)
- Resources must meet methodological and **disciplinary variety**
- Establishing infrastructures requires **resources**
- Resources have to be user friendly | easy to use, and **intuitive**



Infrastructure Projects in Australia

ATAP: The Australian Text Analytics Platform

- Collaborative, cloud-based workbench environment, bringing together users and providers of data and text analytics tools. It will support researchers transitioning to code-based text analysis, with the resultant benefits of flexibility, reproducibility and reuse.

LADAL: Language Technology and Data Analysis Laboratory

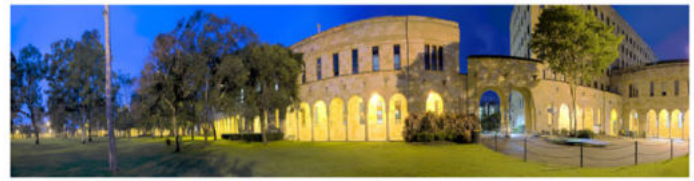
- Free, open-source, collaborative support infrastructure for computational humanities at UQ that offers introductions to topics and concepts related text analytics and practical tutorials, interactive Jupyter notebooks, and events including workshops and webinars.

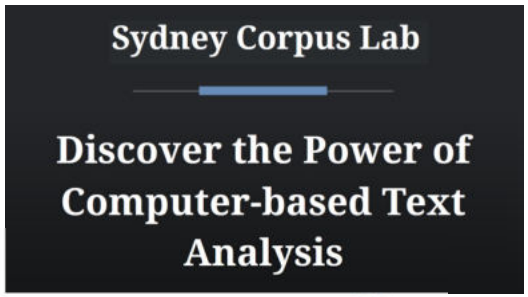
Sydney Corpus Lab

- Promotes corpus linguistics in Australia in linguistics and in other disciplines and aims to build research capacity in corpus linguistics at USydney with strong links to the Sydney Centre for Language Research and the Sydney Digital Humanities Research Group (as well as the Sydney Informatics Hub).

LDaCA: The Language Data Commons of Australia

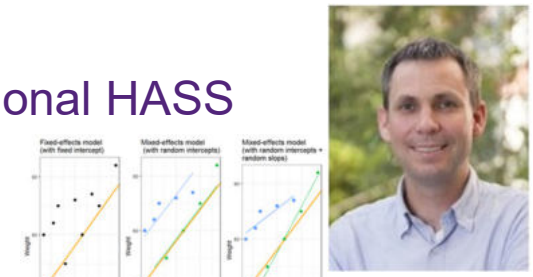
- LDaCA will make nationally significant language data available for academic and non-academic use and provide a model for ensuring continued access with appropriate community control.





Language Technology and Data Analysis Laboratory (LADAL)

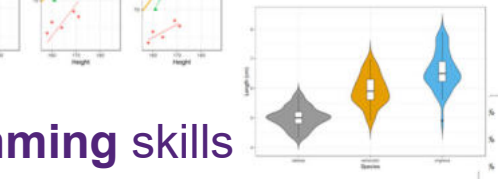
eResearch support infrastructure for computational HASS in the UQ School of Languages and Cultures

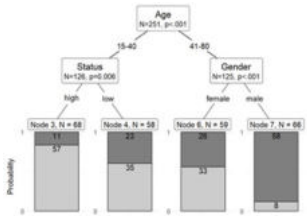


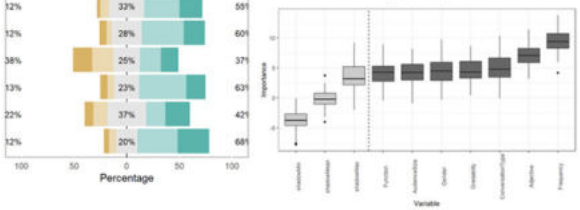
Michael Haugh (co-director of LADAL)

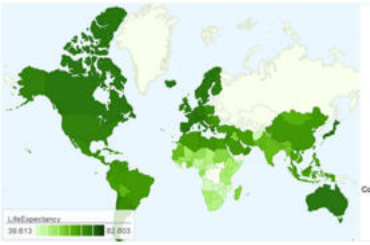
Enables **development** of skills in

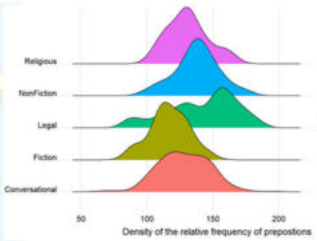
- Digital tools and **data management**
- **Computational methods** and (basic) **programming** skills
- **Data** extraction / transformation / processing
- Data **visualization** (including geospatial mapping and interactive web apps)
- **NLP** applications (text analytics)
- Various statistical procedures (including classification and machine learning)

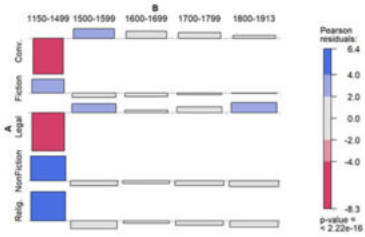








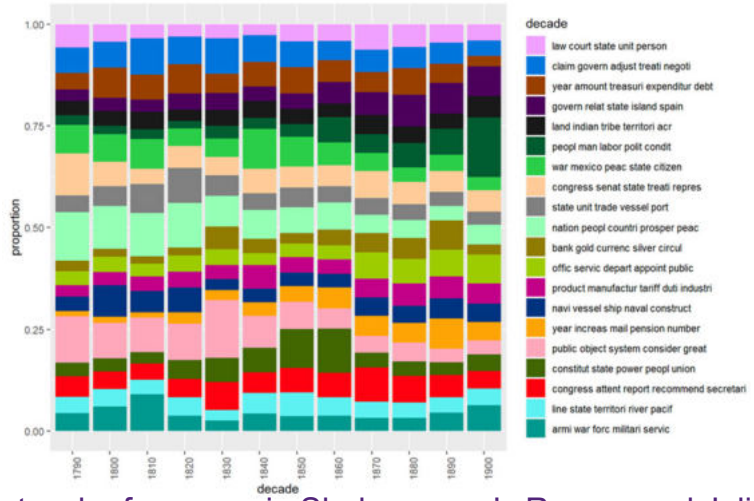




LADAL: https://ladal.edu.au

Language Technology and Data Analysis Laboratory (LADAL)

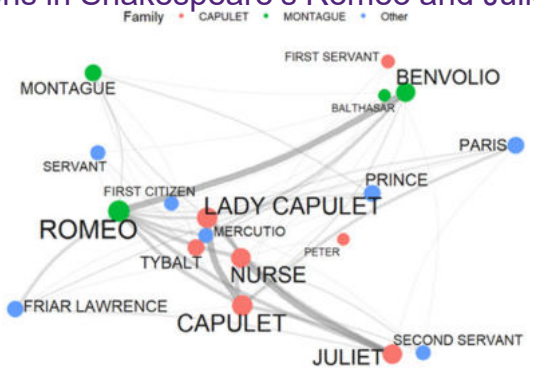
Distribution of topics in US State of the Union Addresses over time



What we hope to achieve

- **Improve transparency and quality** by showcasing how to produce reproducible workflows in R/RStudio)
- Enable researchers to pursue **new pathways** by using innovative methods and new types of data
- Improve **data management,** assist in making workflows **tidier**, more **transparent** and **more efficient**.
- Provide an **infrastructure for acquiring computational skills** (relevant for academia | employability for graduates)
- Showcase how **CL methods more attractive to related disciplines**

Network of persons in Shakespeare's Romeo and Juliet



Problems associated with increased transparency (in CL)

Problems associated with transparency

- Field becomes even **more computational** (shift away from language to technology)
  - Suggestions to learn several programming language (Python, R, Java, and C) are simply unrealistic (and show a disconnect with what HASS researchers/linguists do)
- **Linguistics is not software development**
  - **Linguists have limited resources** and they are interested in language use!
  - Parsimony! What tools give researchers the best bang for their buck!
  - Every tool increases the complexity and requires training: the fewer tools the better! (That is why I promote R and RStudio)
- **Upskilling** required!
- **Infrastructures** required!

Problems associated with transparency

- Transparency can be a **hinderance to careers** (time consuming, scooping, contributions not being valued)
- **More work** (adding to an already excessive workload)
- No guarantee transparency will work and **win back trust**
- Transparency for the **few not the many**?! (Lay audience will have difficulty to understand where to find documents and how to work with them)

Discussion and Outlook

Key points

- Corpus Linguists should think about reproducibility and transparency and about ways to make our research more practically transparent (on an individual, team, and community level)
- There are advantages to making one's work reproducible and transparent (transparent folder structures, documentation, re-use of code, etc.)
- CLs have good reasons **not** to become (too) software focused (and they should not be!)
  - Parsimony of tools: what tools provide a maximum of benefit with a minimum of extra upskilling | adding to the workload
- Necessity for infrastructure, upskilling, and resources
  - Programs and courses on computational tools
- There are serious issues that need to be addressed (career development, scooping, acknowledgement of work)

Discussion and Outlook

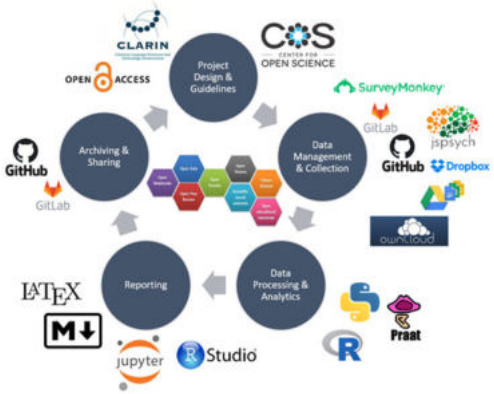
Corpus Linguists are becoming aware of replicability | reproducibility (issues)

- **Arppe et al. (2010)**: “Ideally, research in cognitive linguistics should be based on authentic language use, its results should be **replicable**, and its claims falsifiable. ”
- **Kortmann (2018)**: “do everything that is necessary (!) for achieving a maximum of methodological **transparency**, rigour, statistical significance, **robustness, reproducibility**, falsifiability and, ultimately, explanatory power and mileage for linguistic theory-building”
- Workshops (ISLE5, ISLE6, ICAME42)
- *Linguistics* (upcoming issue)

Discussion on reproducibility and transparency (on a communal and team-level): understanding concepts and aims, integration of tools and methods that make research more transparent

Adopt resources and establish an infrastructure like the infrastructure for quantitative methods (books, workshops, courses, programs, etc.)

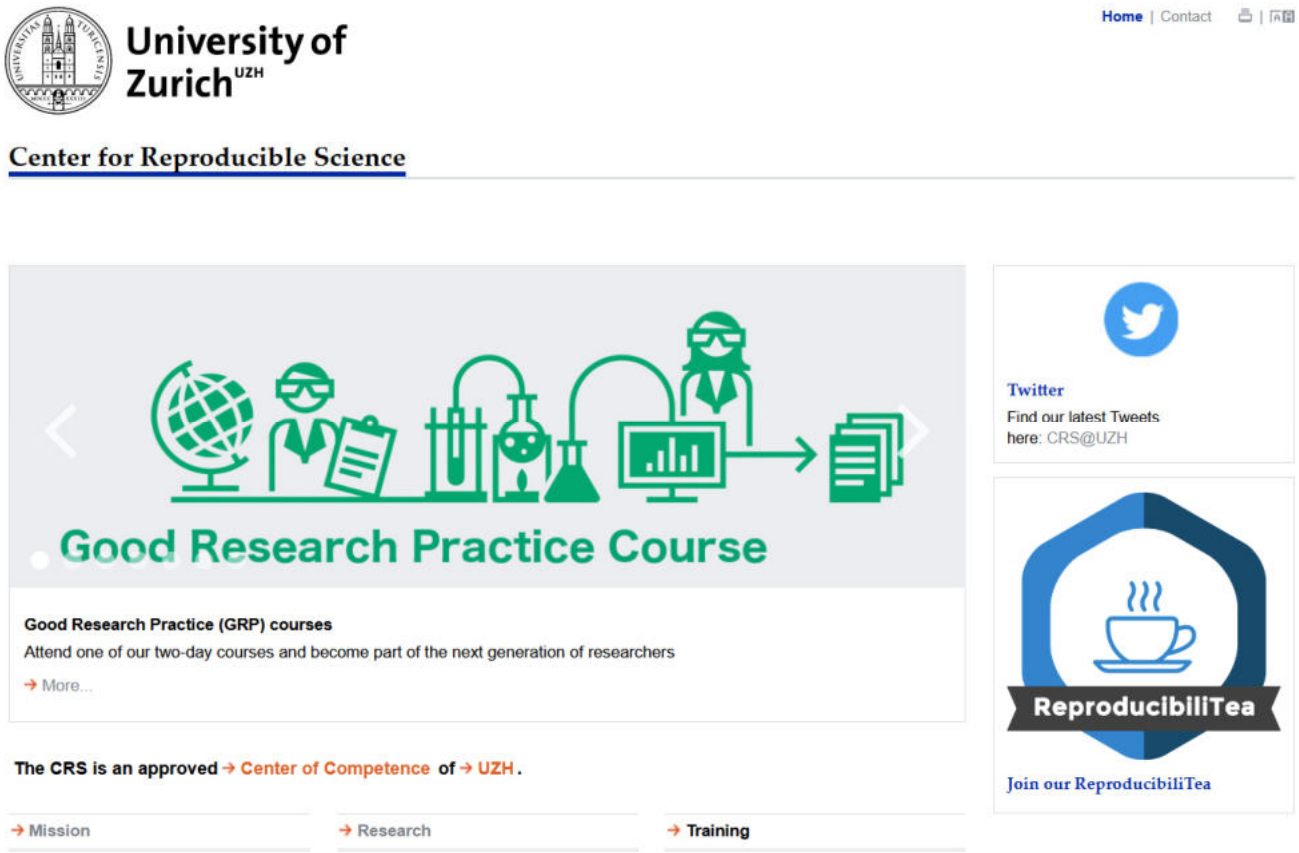
Understanding that it is **NOT (only) a technical (reproduction) issue but a transparency issue**!



Maybe an Australian Center for Reproducible Research?

UZH CRS

The University of Zurich invested in a *Center for Reproducible Science* (https://www.crs.uzh.ch/en.html) to support researchers and develop resources and provide training…



What are your thoughts?

Thank you very much

THE UNIVERSITY OF QUEENSLAND AUSTRALIA
CREATE CHANGE

LADA

NORGES ARKTISKE UNIVERSITET • UIT •

Logosymbol.

AcqVA Aurora Lab

References

Anderson, C. J. et. al. (2016). Response to Comment on “Estimating the reproducibility of psychological science”. *Science* 351(6277): 1037.

Anthony, L. (2020). Programming for Corpus Linguistics. In Magali Paquot & Stefan Th. Gries (eds.), *A practical handbook of corpus linguistics*, 181-207. Berlin & New York: Springer.

Arppe, A., Gilquin, G., Glynn, D., Hilpert, M. & Zeschel, A. 2010. Cognitive Corpus Linguistics: five points of debate on current theory and methodology. *Corpora* 5(1): 1-27.

Baker, M. 2016. 1,500 scientists lift the lid on reproducibility. *Nature* 533: 452-454.

Desagulier, G. (2017). *Corpus linguistics and statistics with R.* Springer International Publishing.

Gilbert, D. T., King, G., Pettigrew, S., Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science". *Science* 351(6277): 1037.

Gries, S. T. (2009). What is corpus linguistics? *Language and Linguistics Compass* 3: 1–17.

Gries, S. T. (2016). *Quantitative corpus linguistics with R: A practical introduction*. Routledge.

References

Gries, S. T. (2021). *Statistics for Linguistics with R*. Berlin: De Gruyter Mouton.

Grieve, J. (2021). Observation, experimentation, and replication in linguistics. *Linguistics.* https://doi-org.ezproxy.library.uq.edu.au/10.1515/ling-2021-0094

Hammond, M. (2020). *Python for Linguists*. Cambridge: Cambridge University Press.

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine* 2(8): e124.

Janda, L A. 2017. The quantitative turn. In B. Dancygier (ed.), *The Cambridge handbook of cognitive linguistics*, 498–514. Cambridge: Cambridge University Press

Kortmann, B. (2018). Reflecting on the quantitative turn in linguistics. Lunch Lecture 2017/18: Quantitative vs. qualitative approaches across sciences (11 January 2018). Accessed Aug. 13, 2021, url: https://www.frias.uni-freiburg.de/downloads/veranstaltungen/ppp-kortmann)

Kortmann, B. (2021). Reflecting on the quantitative turn in linguistics. *Linguistics* https://doi-org.ezproxy.library.uq.edu.au/10.1515/ling-2019-0046

Larsson, T., Egbert, J. & Biber, D. (2022). On the status of statistical reporting versus linguistic description in corpus linguistics: A ten-year perspective. *Corpora* 17(1): 137–157.

References

Levshina, N. (2015). *How to do linguistics with R. Data Exploration and Statistical Analysis*. Amsterdam & Philadelphia: John Benjamins.

Lundquist, H. (2010). *Corpus Linguistics and the Description of English*. Edinburgh: Edinburgh University Press.

McEnery, T. & A. Hardie. (2001). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

Open Science Collaboration,. (2015). Estimating the reproducibility of psychological science. *Science* 349(6251): aac4716.

Sönning, L. & V. Werner . (2021). The replication crisis, scientific revolutions, and linguistics. *Linguistics* 59.5: 1179-1206 (Special Issue: The replication crisis: Implications for linguistics)

References

Australia

LADAL: The Language Technology and Data Analysis Laboratory. School of Languages and Cultures, The University of Queensland. https://ladal.eu.au.

Sydney Corpus Lab: The University of Sydney. https://sydneycorpuslab.com/

ATAP: The Australian Text Analytics Platform. https://www.atap.edu.au.

LDaCA: The Language Data Commons of Australia, LDaCA. https://www.ldaca.edu.au/

Norway

TROLLing: The Tromsø Repository of Language and Linguistics (part of DataverseNO and CLARIN C Centre): https://dataverse.no/dataverse/trolling

AcqVQ Aurora Lab. UiT Aurora Center for Language Acquisition, Variation, and Attrition, The Arctic University of Norway, Tromsø. https://site.uit.no/acqvalab/.

PoLaR: Psycholinguistics of Language Representation (PoLaR) lab. UiT Aurora Center for Language Acquisition, Variation, and Attrition, The Arctic University of Norway, Tromsø. https://site.uit.no/polar/.

Switzerland

UZH-CRS: Center for Reproducible Science, University of Zurich. https://www.crs.uzh.ch/en.html

slides available at

www.martinschweinberger.de

THE UNIVERSITY OF QUEENSLAND AUSTRALIA
CREATE CHANGE

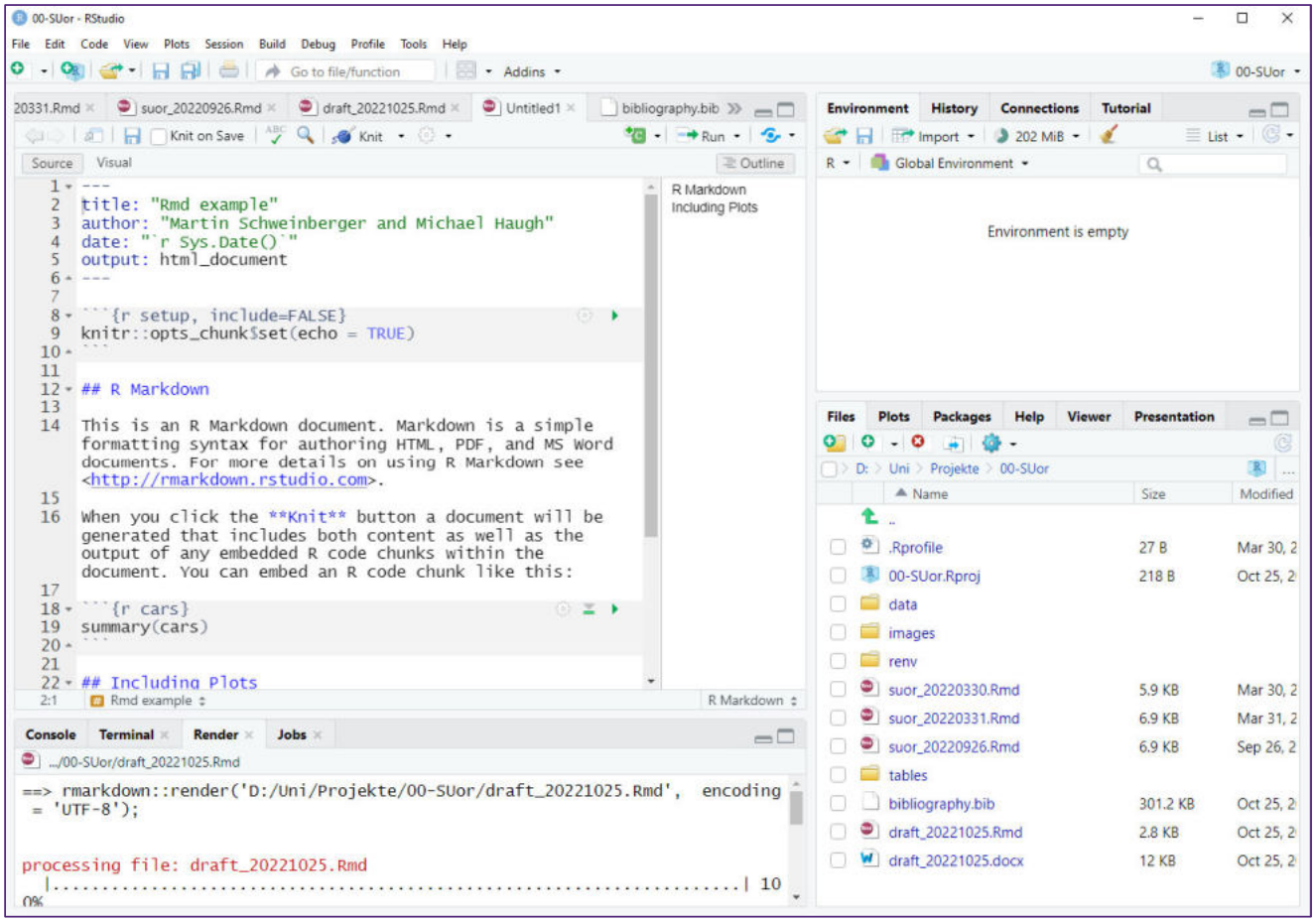
LADAL

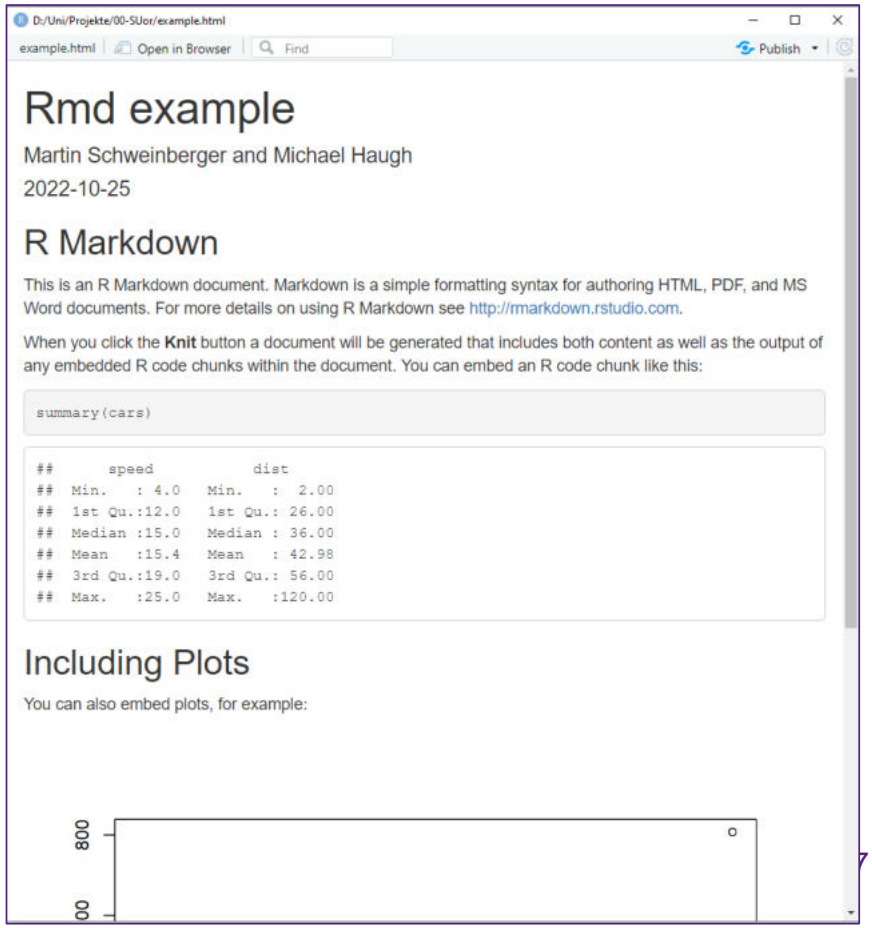
NORGES ARKTISKE UNIVERSITET • UIT •

Logosymbol.

AcqVA Aurora Lab

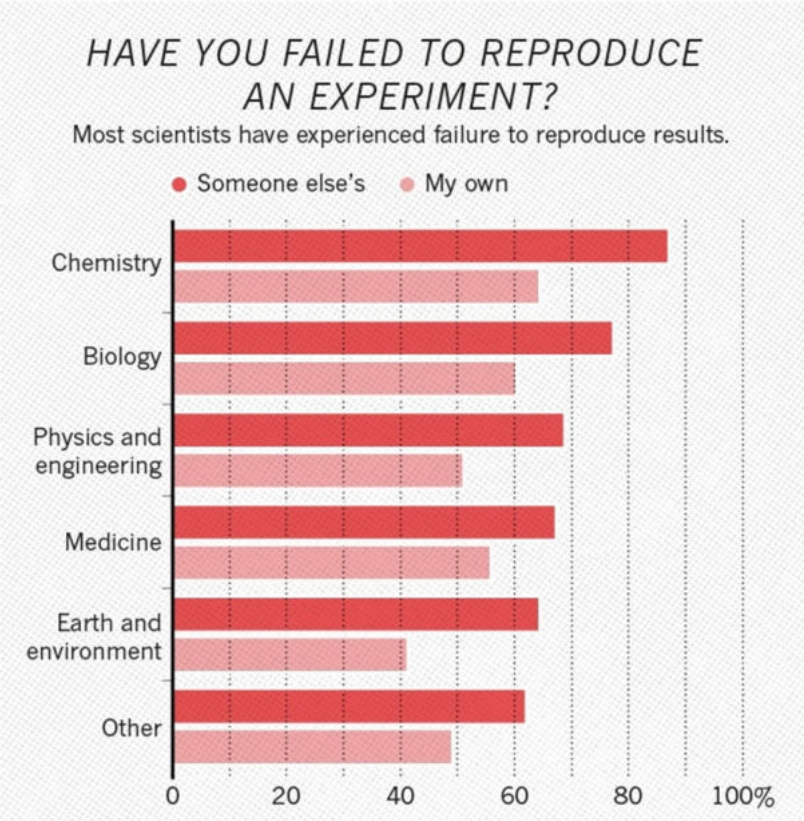
Example of a RNotebook (raw and rendered)



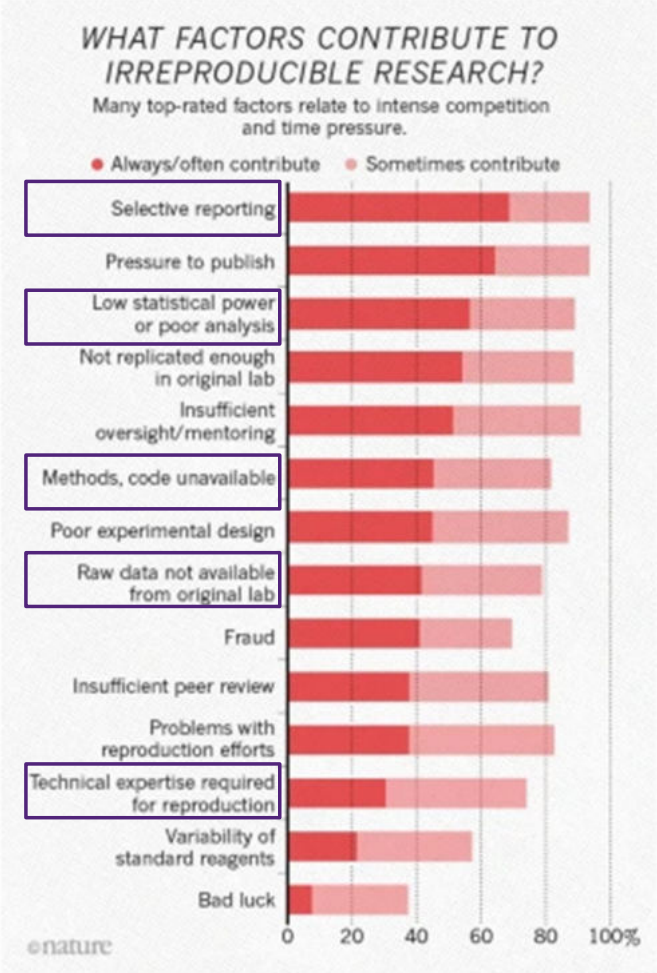


Replication Crisis | Issue | Problem

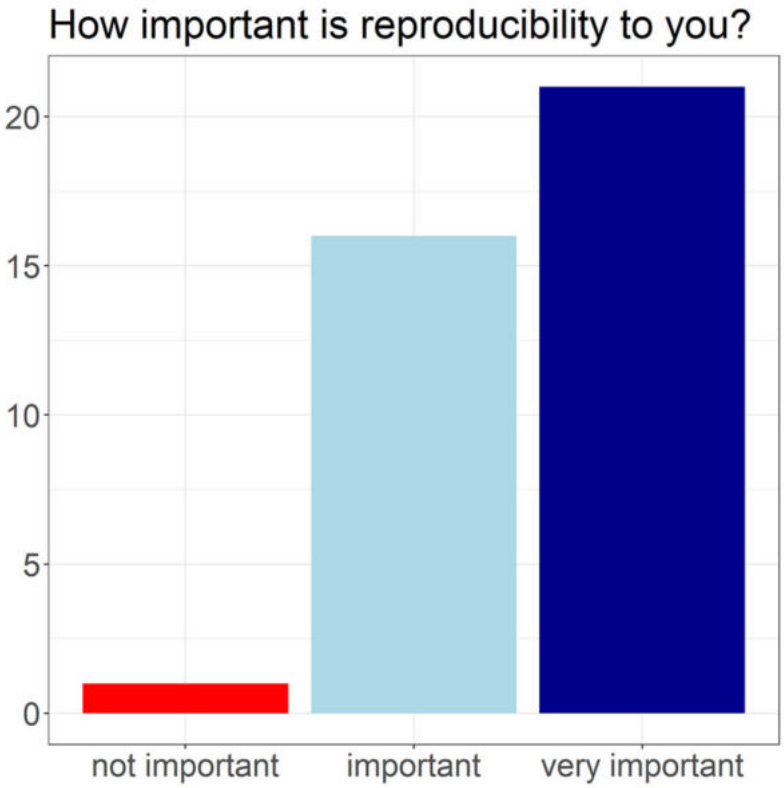
from Baker (2016: 452)

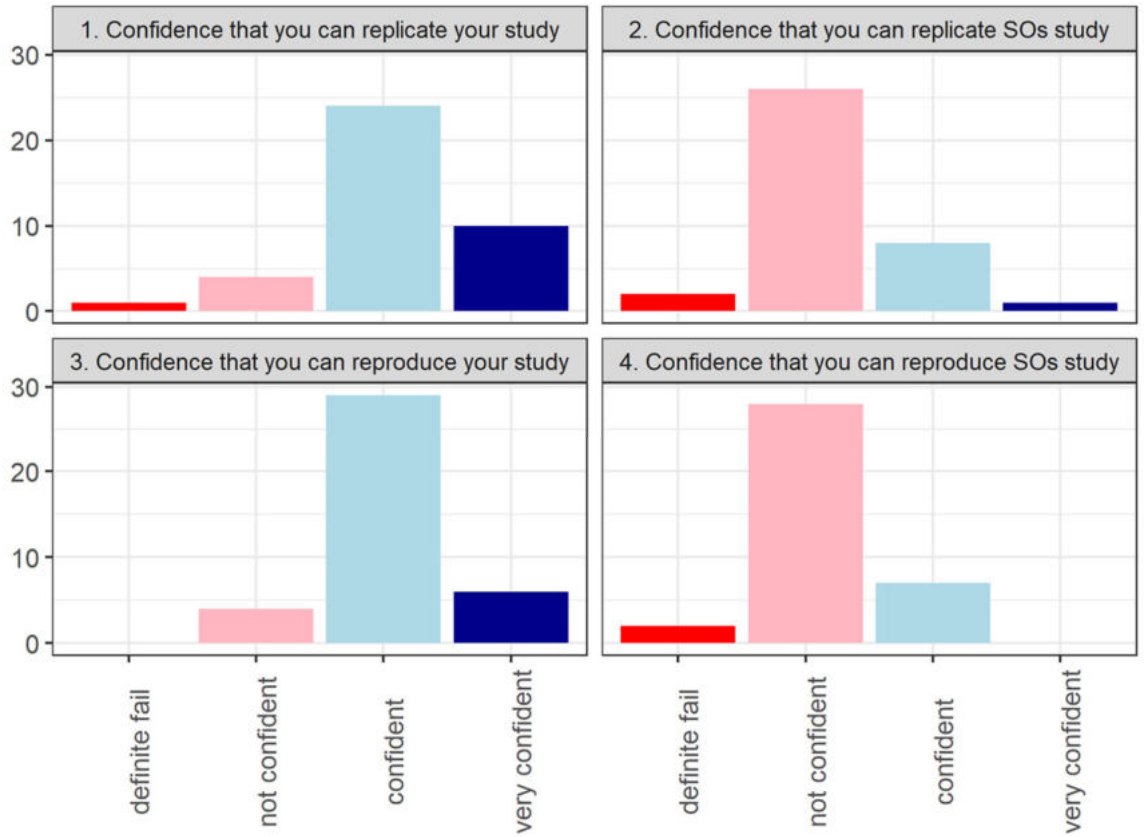


from Baker (2016: 452)



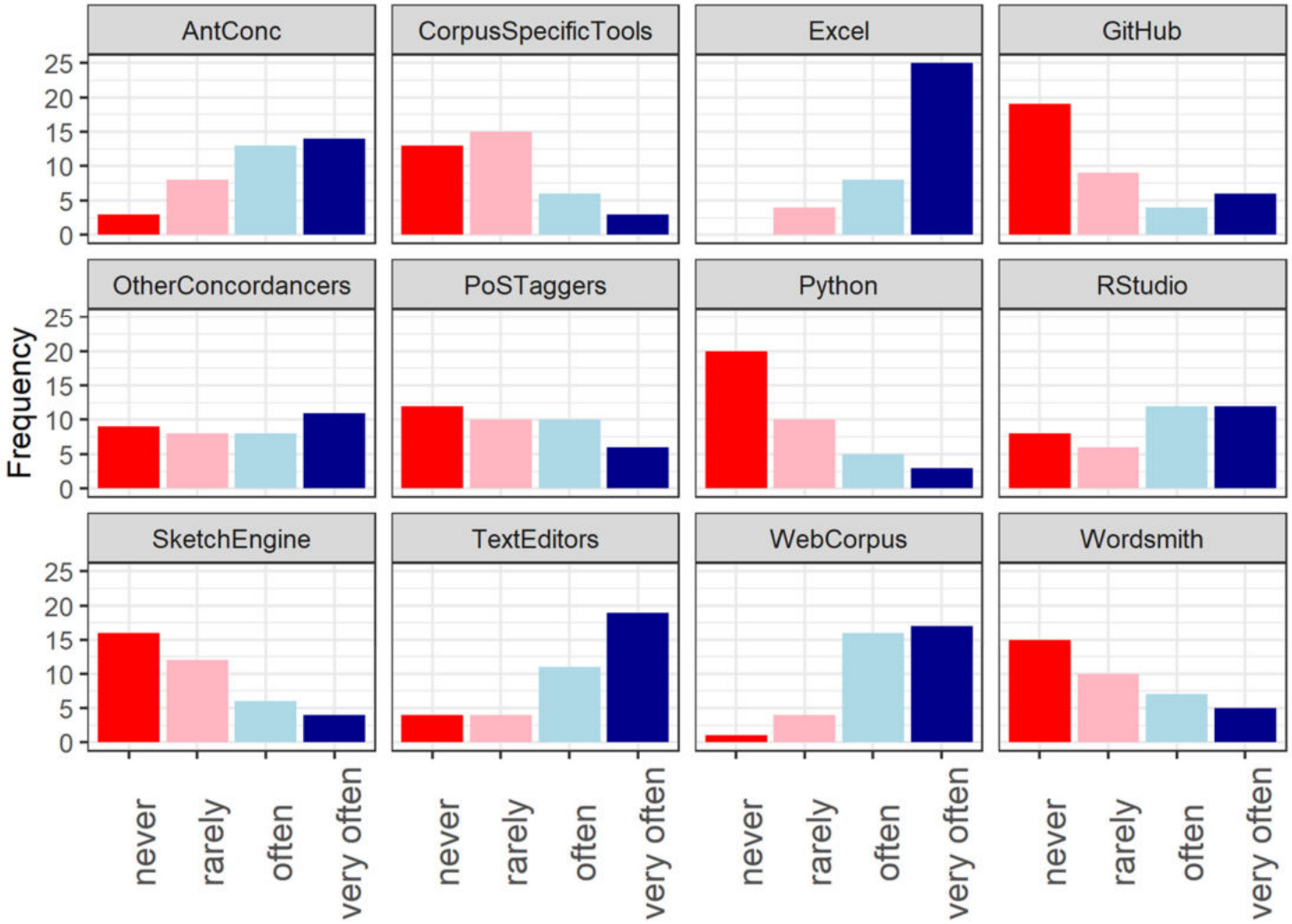
Replication Survey (at ICAME42, 2021, N: 48)





Broad support and acknowledgement that reproducibility is important

We trust ourselves but not other (others **don't TRUST** us)


AntConc
CorpusSpecificTools
Excel
GitHub
OtherConcordancers
PoSTaggers
Python
RStudio
SketchEngine
TextEditors
WebCorpus
Wordsmith
Frequency
0
5
10
15
20
25
never
rarely
often
very often