



Replication and Reproducibility in Social Sciences and Statistics: Context, Concerns, and Concrete Measures

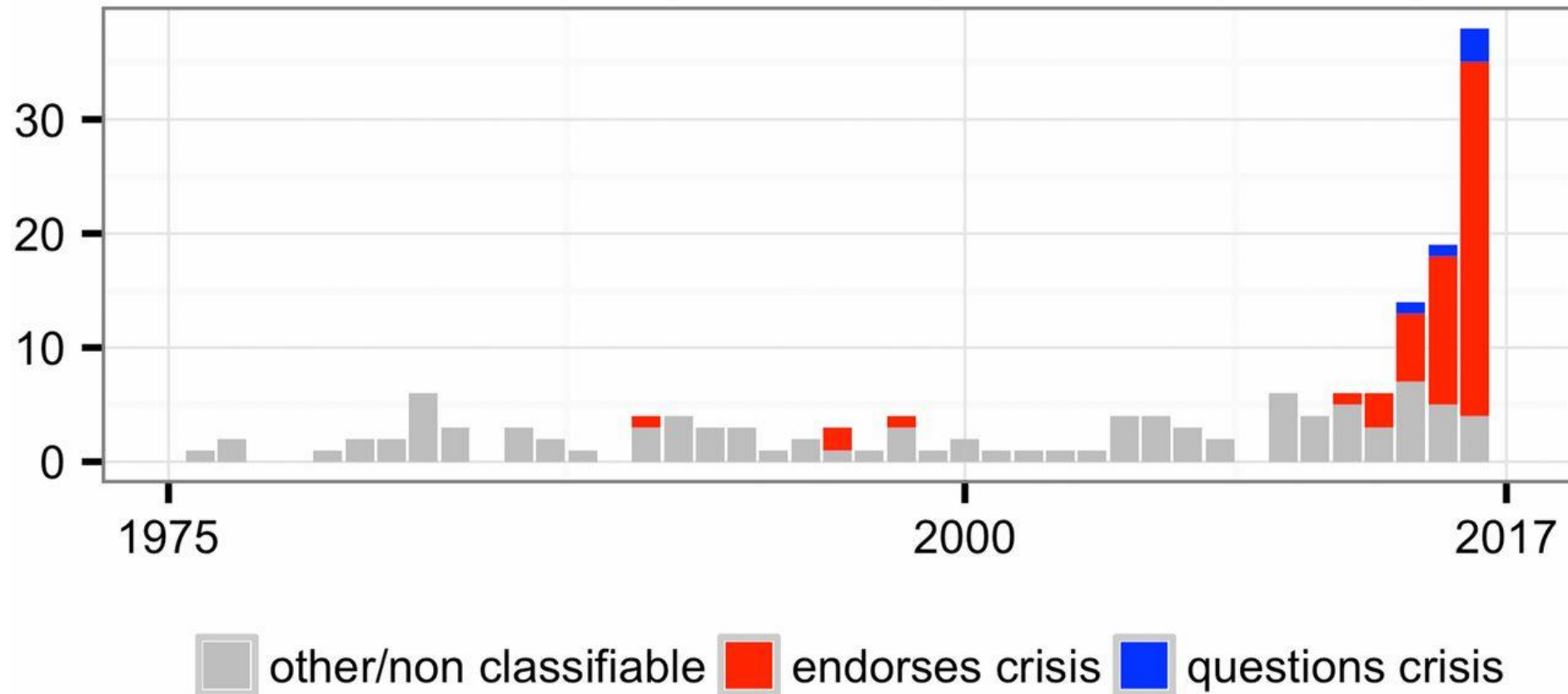
Lars Vilhuber
Cornell University

Partial funding acknowledged under NSF-[#1131848 \(NCRN\)](#) and a grant from the Alfred P. Sloan Foundation.
The opinions expressed in this talk are solely the authors, and do not represent the views of the U.S. Census Bureau, the American Economic Association, or any of the funding agencies.



This reproducibility crisis thing....

Frequency of Crisis Narrative in Web of Science Records





The “crisis” in the 60s and 70s

Sterling, 1959; Cohen, 1962; Lykken, 1968; Tukey, 1969;
Greenwald, 1975; Meehl, 1978; Rosenthal, 1979

Low power

Flexibility in analysis

Selective reporting

Ignoring nulls

Lack of replication

Misuse of statistics

Source: Nosek
Sackler talk 2017



Efficiency of scholarly discourse?

- Early publications (20th century) contained **tables of data**, and the **math** was simple (maybe)
 - **Data** became electronic, was no longer **included** or **cited**
 - **Math** was transcribed to **code**, and was no longer **included**



Efficiency of scholarly discourse!

**Modern publications thus need
the same transparency and completeness
as in the old days
to facilitate replicability**

Progress



Progress

- Replication archives and Data (Code) Availability policies





Progress


- Replication archives and Data (Code) Availability policies
- Shared open source software



Statistical Software Components

From [Boston College Department of Economics](#)
Boston College, 140 Commonwealth Avenue, Chestnut Hill MA 02467 U:
Contact information at [EDIRC](#).
Bibliographic data for series maintained by Christopher F Baum ([baum@](#)

[Access Statistics](#) for this software series.
Track citations for all items by [RSS feed](#)
Is something missing from the series or not right? See the RePEc data [series](#).

[GAPPORT: Stata module to calculates seats in party-list representation](#)  downloads
Ulrich Kohler

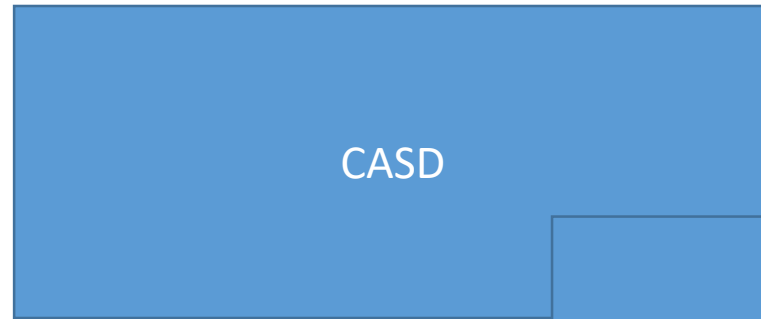
[GCLSORT: Stata module to sort a single variable via egen](#)
Philippe Van Kerm

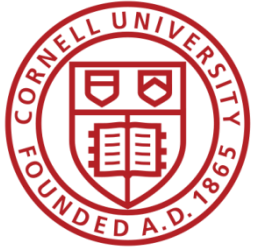
[GPROD: Stata module to extend egen for product of obs](#)
Philip Ryan



Progress

- Replication archives and Data (Code) Availability policies
- Shared open source software
- Better public-use and shared data





Progress

- Replication archives and Data (Code) Availability policies
- Shared open source software
- Better public-use and shared data
- Better ways of accessing preprints/ grey literature

RePEc



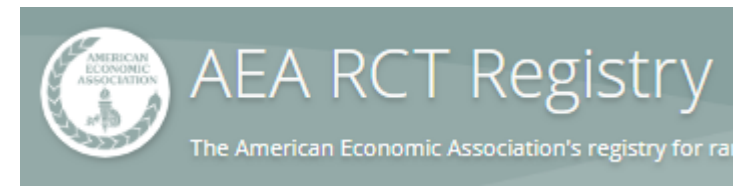


Progress

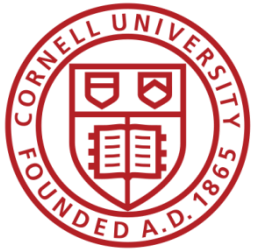
- Replication archives and Data (Code) Availability policies
- Shared open source software
- Better public-use and shared data
- Better ways of accessing preprints/ grey literature
- Pre-registration of trials, experiments, and analyses

It's registered

Osf.io



More
recently...



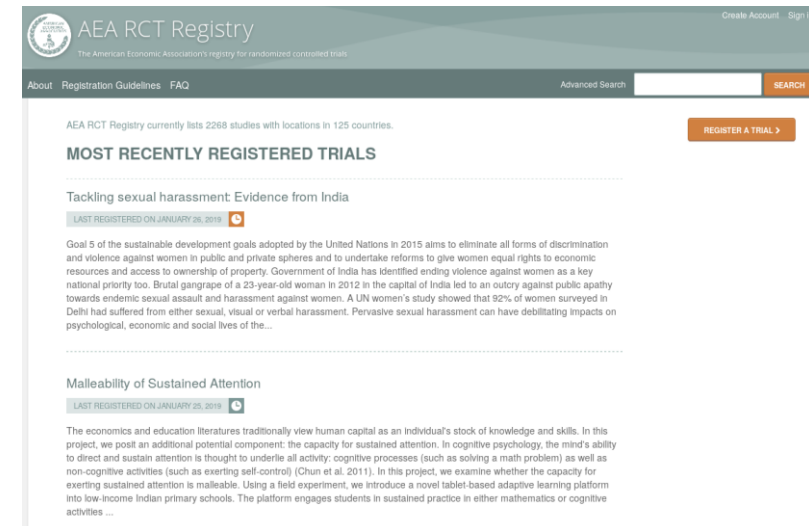
Second round (2012-)

- **Greater enforcement of data (and code) availability**
 - 2015, AJ Political Science
 - 2016, Data Editor for ASA Software Section
 - 2016, Statistical review added Science
 - 2017: AEA appoints Data Editor, with mandate to do similar activities



Pre-registration

- “That information is especially helpful in research that emphasizes **null hypothesis significance testing**.
- A thorough preregistration promotes transparency and openness and **protects researchers from suspicions of p-hacking.**”





Registered Reports

- <https://cos.io/rr>
- Chambers (2014)
- Nosek & Lakens (2014)



- Close cousin: Results-blind review



Preprints in other sciences

- bioRxiv (2013)
- PsyArXiv (2016)

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

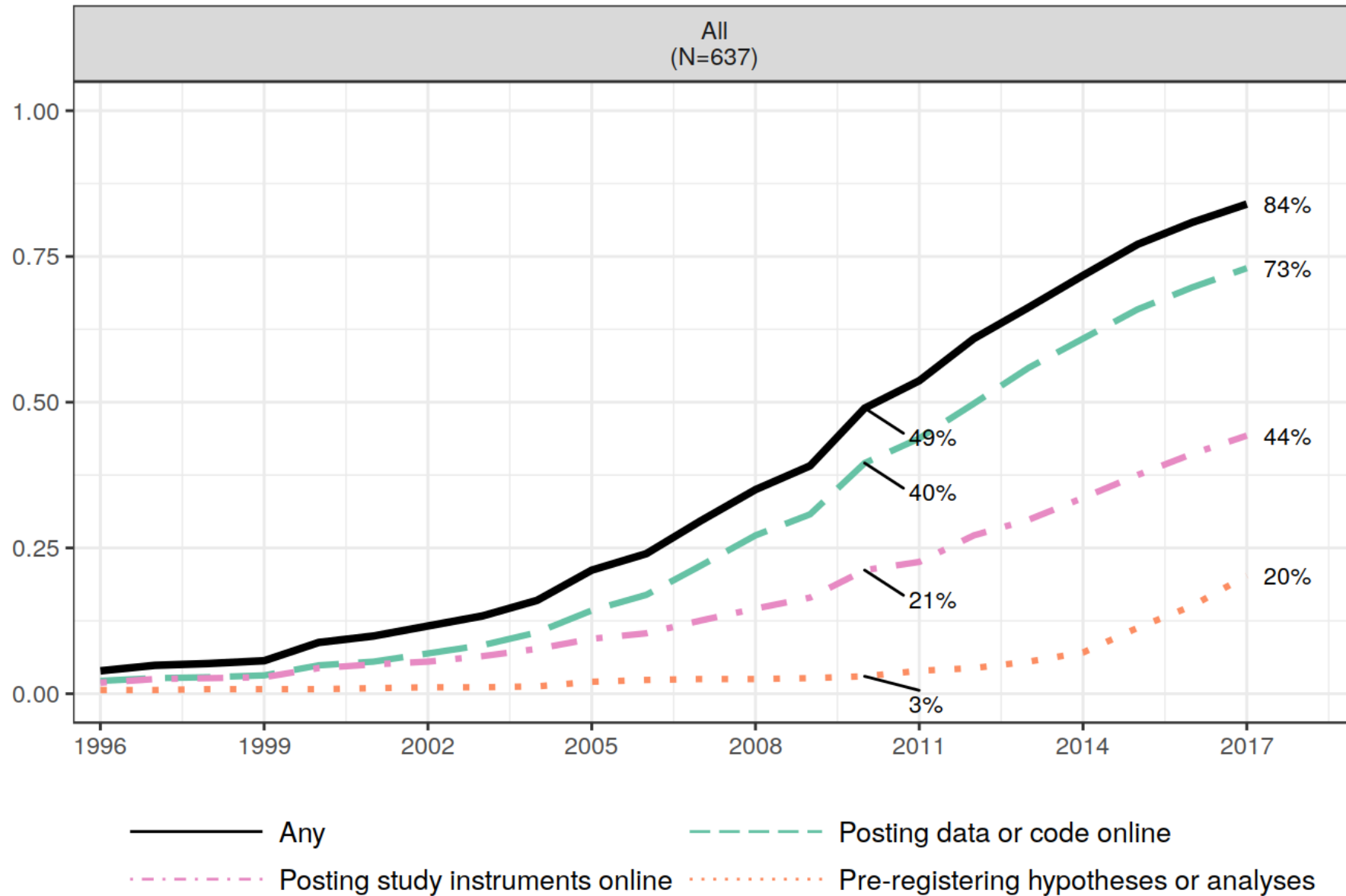
Search

[Advanced Search](#)





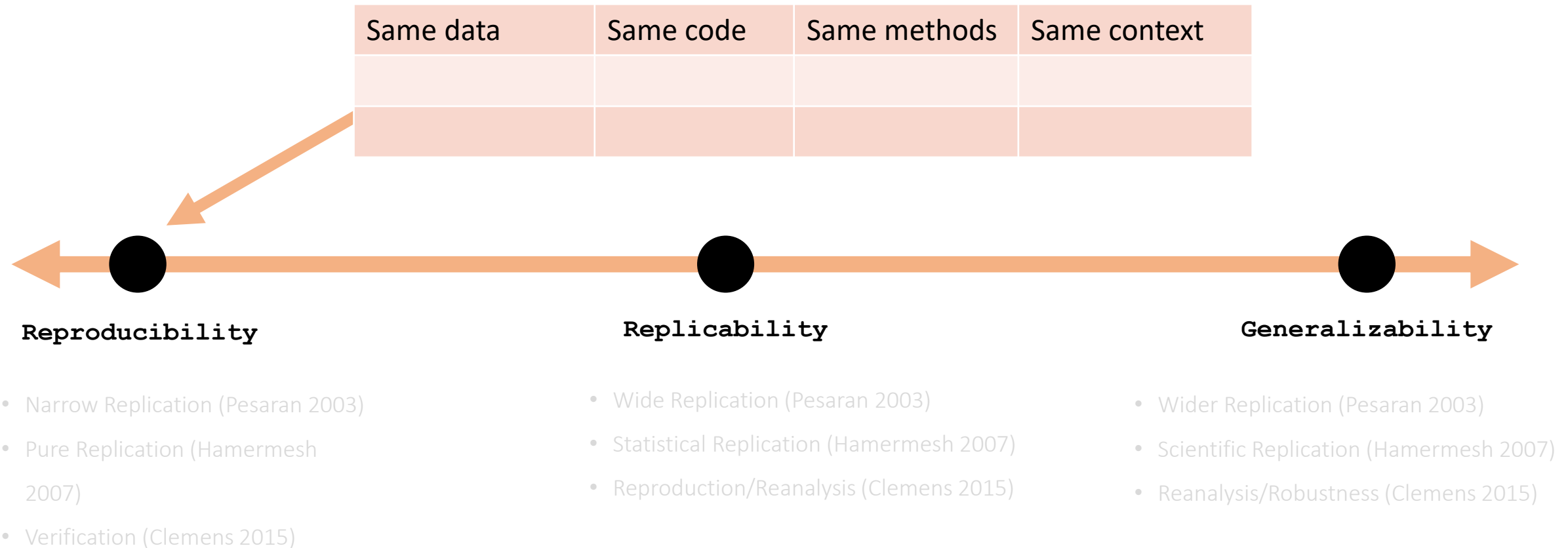
Share of Published Authors (PhD < 2010) Adopting Practice



Issues

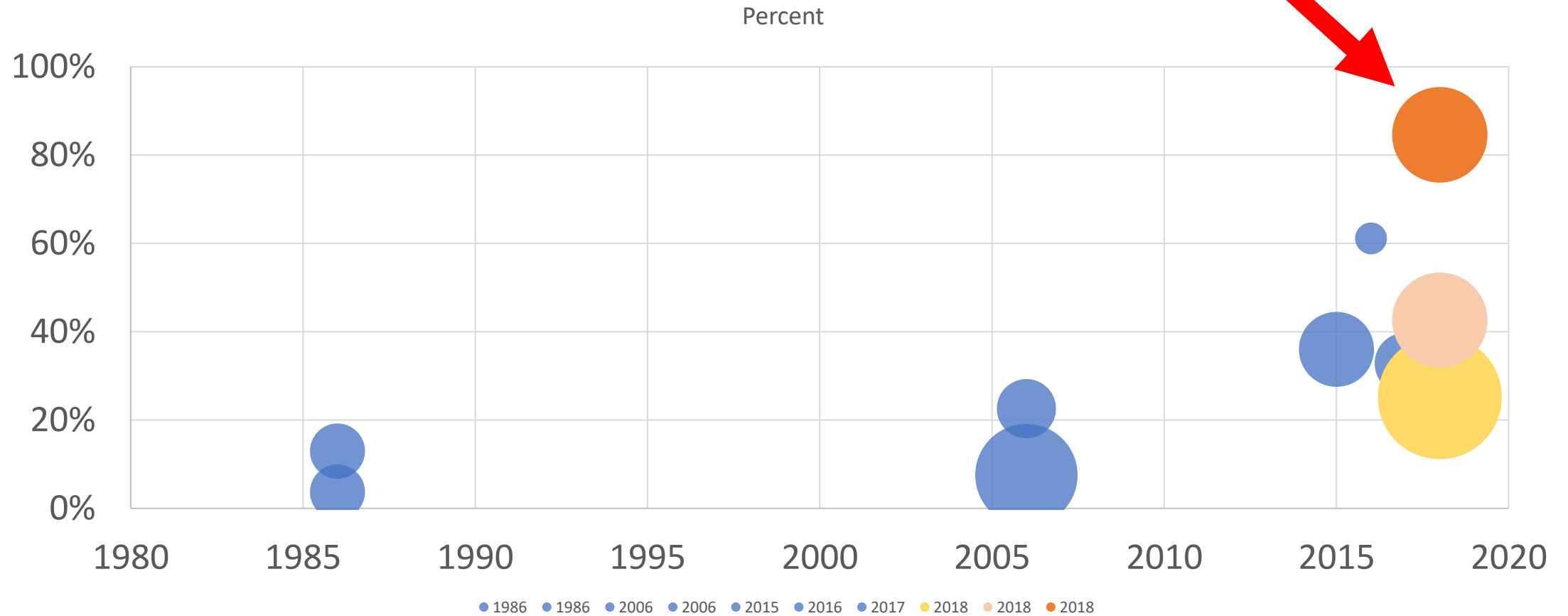


Replication continuum





Results?





Some key statistics

Study	Year	N	Success	Type	Type-R	Type-Data	Percent	Field
Dewald Thursby Anderson	1986	54	2	Complete	Reproducibility	Avail	4%	Economics
Dewald Thursby Anderson	1986	54	7	Partial	Reproducibility	Avail	13%	Economics
McCullough McGeary Harrison	2006	186	14	Complete	Reproducibility	All	8%	Economics
McCullough McGeary Harrison	2006	62	14	Complete	Reproducibility	Avail	23%	Economics
Nosek et al	2015	100	36	Complete	Replication		36%	Psychology
Camerer et al	2016	18	11	Complete	Replication		61%	Experimental Econ
Chang Li	2017	67	22				33%	Macroeconomics
Kingi et al	2018	274	69	Complete	Reproducibility	All	25%	Economics
Kingi et al	2018	162	69	Complete	Reproducibility	Avail	43%	Economics
Kingi et al	2018	162	137	Partial	Reproducibility	Avail	85%	Economics

Kingi et al numbers are preliminary. Do not cite or quote.

Not enough
articles are
reproducible



Current Data Availability Policies are Broken

- If the Data is
not open-access,

**no systematic information is
collected**
(“exemption”)

Not enough
data is
“accessible”



Current efforts at the AEA



Current efforts at the AEA

- Provide more transparency
 - To assist replication efforts
 - By better linking to paper-related resources (data, code, registration, etc.)
- Pre-emptively improve code archives
 - By conducting reproducibility checks
 - By working with groups that conduct reproducibility checks
- Better archives
 - Greater transparency of the code and data archives



AEA “Data Availability Policy” (2018)

- It is the policy of the American Economic Association to publish papers only if the data used in the analysis are clearly and precisely documented and are readily available to any researcher for purposes of replication.
- Authors of accepted papers that contain empirical work, simulations, or experimental work must **provide**, prior to publication, the **data, programs, and other details of the computations sufficient to permit replication**. These will be posted on the AEA website. The Editor should be notified at the time of submission if the data used in a paper are proprietary or if, for some other reason, the requirements above cannot be met.

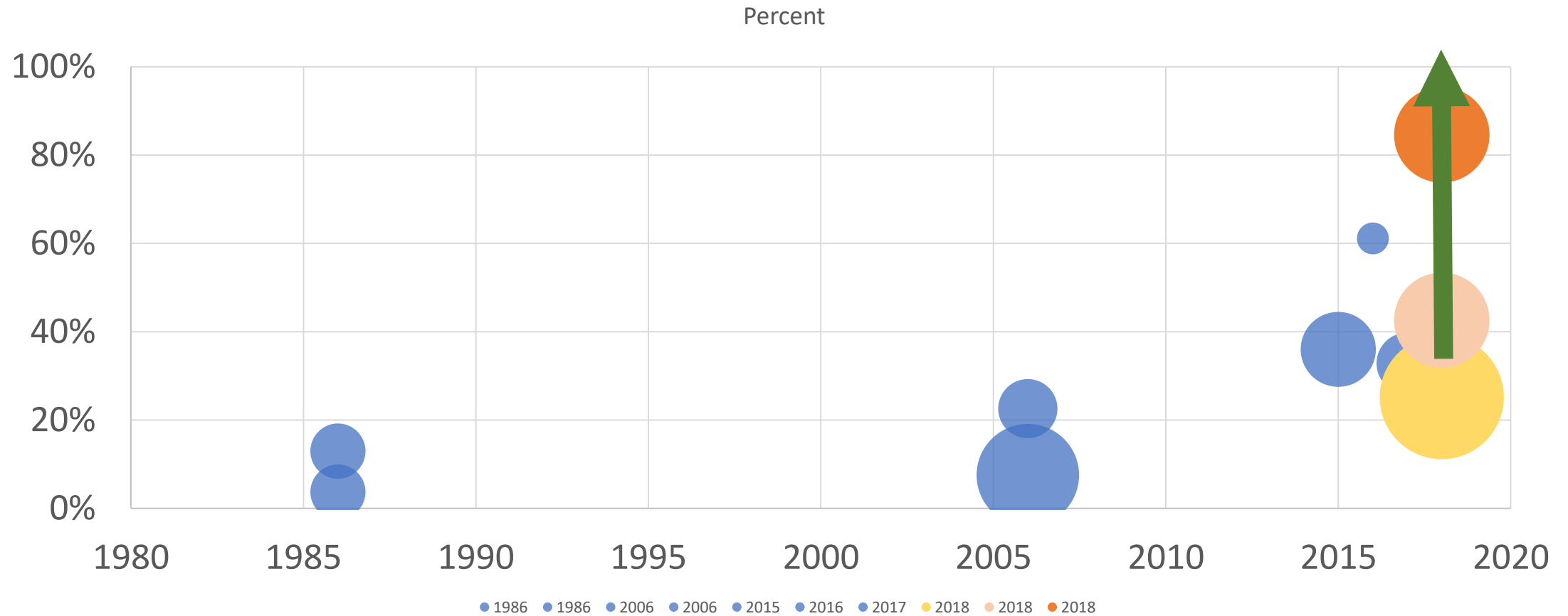


AEA “Data Availability Policy” (2019)

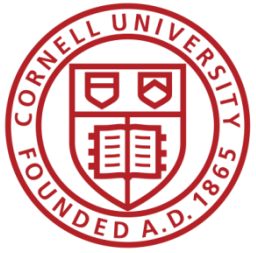
- It is the policy of the American Economic Association to publish papers only if the data used in the analysis are **clearly and precisely** documented and are **readily available** to any researcher for purposes of replication.
- Authors of accepted papers that contain empirical work, simulations, or experimental work must **provide, prior to publication**, the data, programs, and other details of the computations **sufficient to permit replication**. These will be **posted on the AEA website**. The Editor should be notified at the time of submission if the data used in a paper are proprietary or if, for some other reason, the requirements above cannot be met.



Improve reproducibility



It is not the
access that is
“broken”



Illustration

If you used files at
the National Archives,

would we ask you to
“deposit” them?

It is the
description of
access that is
“broken”



Encourage Best Practices

- **Deposit and archive early**
 - If you collect data, archive it
(possibly privately)
 - If you finish the manuscript,
deposit the analysis files
(possibly privately)



Encourage Best Practices

- **Follow robust coding**
 - Ensure that code reliably produces results
(possibly automated)
 - Before you finish the manuscript, run all analysis code again
(if not too onerous)



Evolving Journal and Data Infrastructure

- More self-deposit repositories in the social sciences
 - Dataverse
 - Figshare
 - openICPSR
 - Zenodo
 - Qualitative Data Repository (QDR)
 - Others...



Evolving Journal and Data Infrastructure

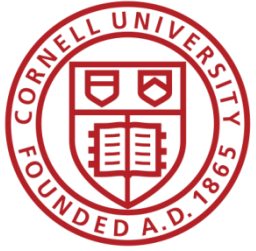
Use them!



Evolving Journal and Data Infrastructure

Describe them!

(cite them!)



Evolving Journal and Data Infrastructure

Treat all archives
symmetrically!



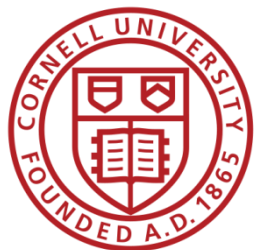
Evolving Journal and Data Infrastructure

- More ~~self-deposit~~ repositories in the social sciences

- Dataverse
- Figshare
- openICPSR
- Zenodo

- **CASD**
- **IAB**
- **Norway**
- **US Federal Statistical RDC**
-

- Qualitative Data Repository (QDR)
- Others...



Challenges?



Verifying Data and Code Deposits

Why do journals like
affiliated repositories
(or website deposits)?

- They can ensure **longevity/ persistence**
- They can ensure **access**
- They can ensure **availability**



Verifying Data and Code Deposits

- Not every data repository is created equal
 - Github, Dropbox, etc. are not data or code repositories
 - Is the institutional repository at the University of Southern Venezuela a reliable repository?
 - Is the institutional repository at Cornell University a reliable repository?
 - Is the institutional repository at Harvard University (Dataverse!) a reliable repository?
 - Are the National Archives a reliable repository?



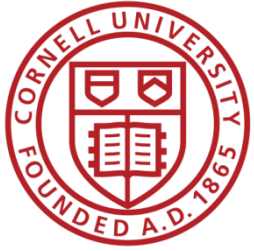
Verifying Data and Code Deposits

- Not every data repository is created equal
 - The Second Bank of Third City credit card data is not a data/code repository
 - Is the School Board of Third City a reliable repository?
 - Is the JPMC Institute a reliable repository?
 - Is the US Census Bureau a reliable repository?
 - **Are any restricted-access repositories reliable archives?**



Within Economics (AEA, Restud)

- Ensure **reproducibility** of computational code
- Challenge: **Restricted-access data**



Future efforts

AEA, Social Sciences, elsewhere



Better support for researchers

- Training in methods (with various centers, institutions, etc.)
 - For current researchers
 - For integration into curriculums
- Tools to streamline the process
 - A few technical things (not described here)
 - Coordinate among journals (no duplicate effort)
- Awareness
 - Consider badges/ certification
 - Address issues with confidential data



Confidential data

- Highlight where confidential data already require replicability
 - IAB
 - Remote processing servers (Canada, NCHS, Australia, etc.)
- Work with Research Data Centers to facilitate transparency and reproducibility
 - Training (secure programming guidelines)
 - Standardize archives within RDCs + transparency



Abstract



References



Online appendix



Supplementary
materials



Notes

Supplementary materials

- Code and Data

Besley, Timothy, and Hannes Mueller. 2018. "Replication data for: Predation, Protection, and Productivity: A Firm-Level Perspective." *American Economic Journal: Macroeconomics*, 10 (2): 184-221. DOI: [10.1257/mac.20160120.data](https://doi.org/10.1257/mac.20160120.data)

cite! ▼

- ☒ Data is freely accessible at [10.1257/mac.20160120.data](https://doi.org/10.1257/mac.20160120.data) under CC BY-NC 4.0.
- ☒ Code verified under AEA guidelines 2.0

- Data

Statistics Norway. 2015. "Firm-level statistics 1975-2013 [dataset]" Norwegian Data Archive [curator], v2. DOI: [10.7654/nda::7643A::34](https://doi.org/10.7654/nda::7643A::34)

cite! ▼

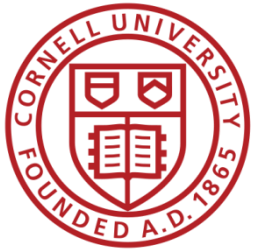
- Data restricted-access (has residency requirement, has citizenship requirement), accessible at Norwegian Data Archive in Oslo, Norway, under Norwegian Data Access license.
- Code could not be verified due to access restrictions.

And you?



Researchers: New skills to learn/teach

- How to **incorporate reproducible practices** into your workflow
- When to **pre-register**, and when not to
- **Document** early, and often (better READMEs!)
- How, where, and when to **archive data and code**
- How to **license** your contributions!



Summary

- **Greater transparency**
 - Equal treatment of public-use and confidential data
- **Better computational reproducibility**
 - For public data as well as confidential data
- **Greater reliance on shared resources**
 - Encourage best practices

Merci!