

Replication and Reproducibility in Social Sciences and Statistics: Context, Concerns, and Concrete Measures **when Data are Confidential**

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.



Disclaimer

My opinions

Not Census Bureau

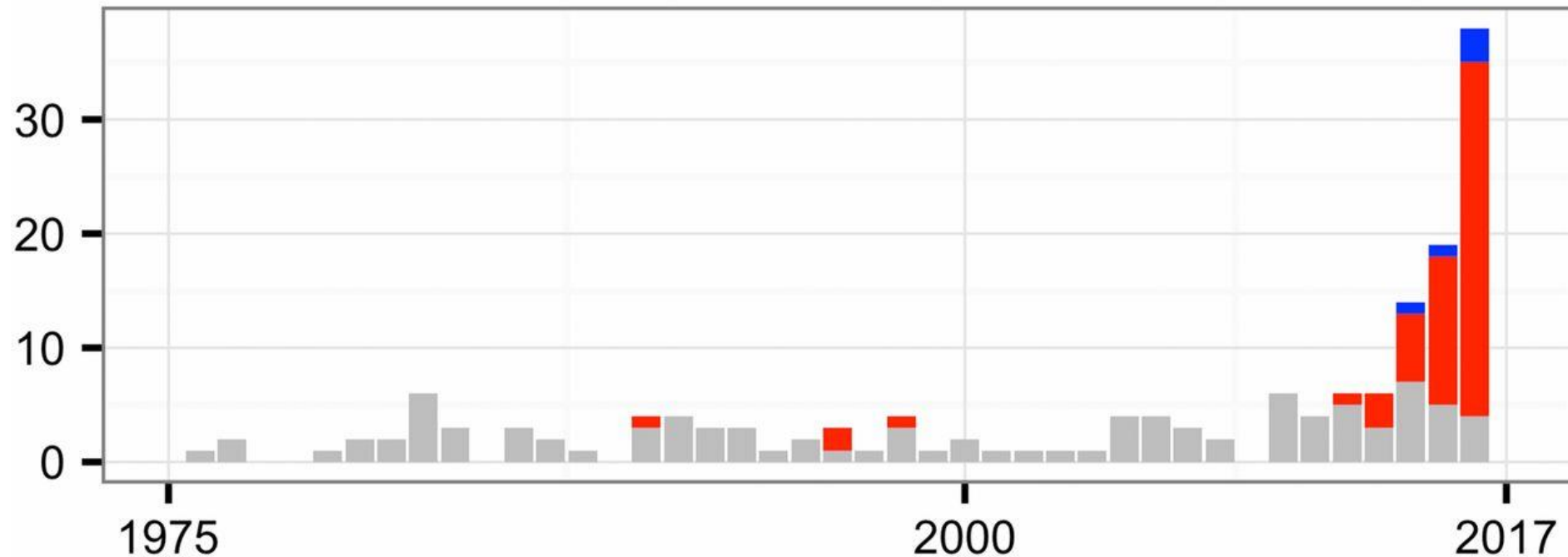
Not American Economic Association

No confidential data was abused in this talk



This reproducibility crisis thing....

Frequency of Crisis Narrative in Web of Science Records



other/non classifiable endorses crisis questions crisis

<https://doi.org/10.1073/pnas.1708272114>



The “crisis” in the 60s and 70s

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

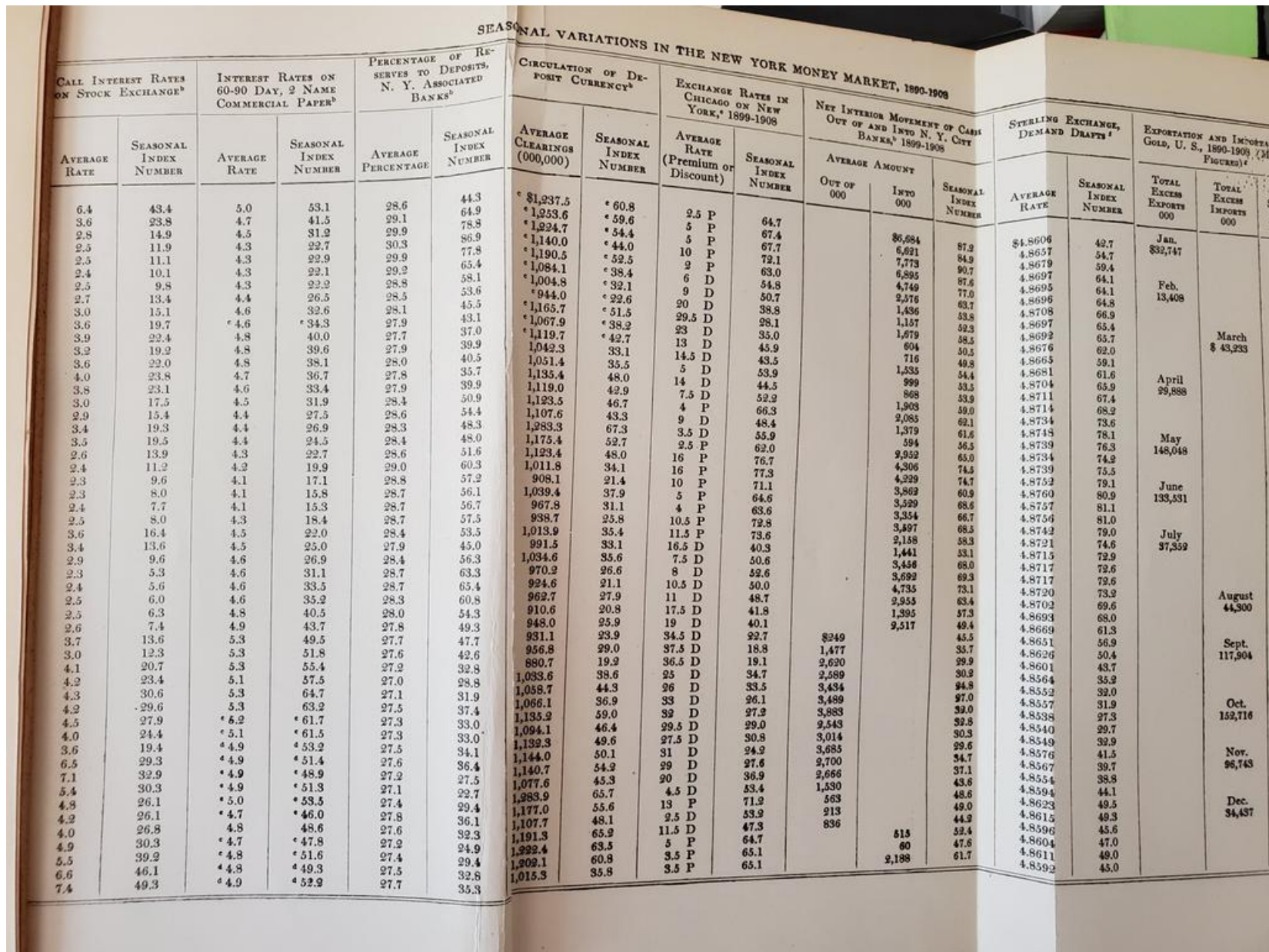
Low power
Flexibility in analysis
Selective reporting
Ignoring nulls
Lack of replication
Misuse of statistics

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**



From @sdellavi
AER 1911



Efficiency of scholarly discourse!

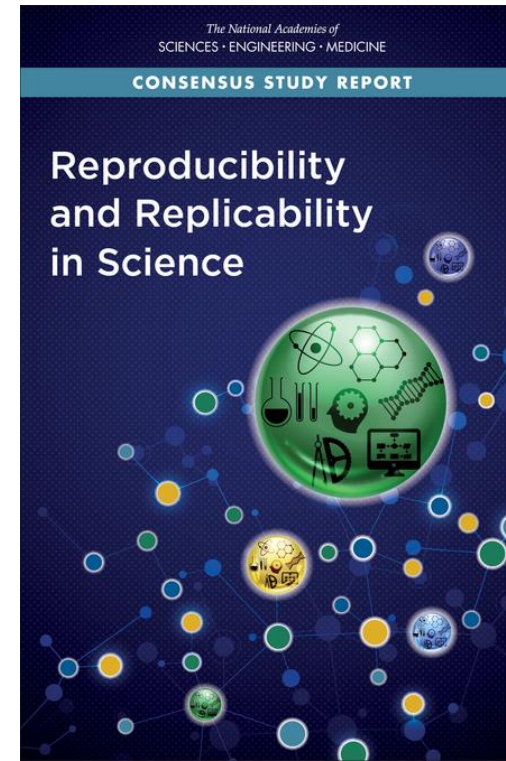
**Modern publications thus need
the same transparency and completeness
as in the old days
to facilitate replicability**

Replication?



Replication continuum

<https://doi.org/10.17226/25303>

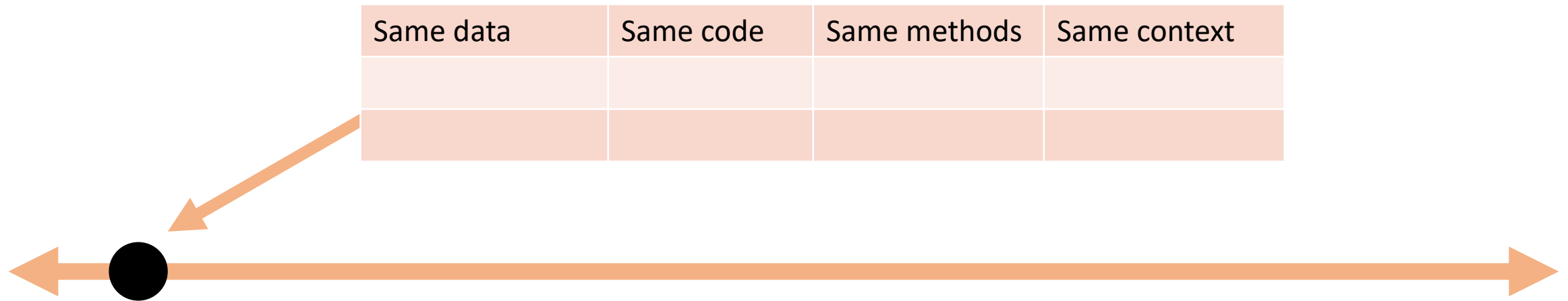


Reproducibility

- Narrow Replication (Pesaran 2003)
- Pure Replication (Hamermesh 2007)
- Verification (Clemens 2015)



Replication continuum



Reproducibility

- Narrow Replication (Pesaran 2003)
- Pure Replication (Hamermesh 2007)
- Verification (Clemens 2015)



Replication continuum



Reproducibility

- Narrow Replication (Pesaran 2003)
- Pure Replication (Hamermesh 2007)
- Verification (Clemens 2015)

Replicability

- Wide Replication (Pesaran 2003)
- Statistical Replication (Hamermesh 2007)
- Reproduction/Reanalysis (Clemens 2015)



Replication continuum

Same data	Different code or software	Same methods	Same context



Reproducibility

Replicability

- Narrow Replication (Pesaran 2003)
- Pure Replication (Hamermesh 2007)
- Verification (Clemens 2015)

- Wide Replication (Pesaran 2003)
- Statistical Replication (Hamermesh 2007)
- Production/Reanalysis (Clemens 2015)



Replication continuum

New data collection	Same code	Same methods	Same context



Reproducibility

Replicability

- Narrow Replication (Pesaran 2003)
- Pure Replication (Hamermesh 2007)
- Verification (Clemens 2015)

- Wide Replication (Pesaran 2003)
- Statistical Replication (Hamermesh 2007)
- Reproduction/Reanalysis (Clemens 2015)



Replication continuum



Reproducibility

- Narrow Replication (Pesaran 2003)
- Pure Replication (Hamermesh 2007)
- Verification (Clemens 2015)

Replicability

- Wide Replication (Pesaran 2003)
- Statistical Replication (Hamermesh 2007)
- Reproduction/Reanalysis (Clemens 2015)

Generalizability

- Wider Replication (Pesaran 2003)
- Scientific Replication (Hamermesh 2007)
- Reanalysis/Robustness (Clemens 2015)



Replication continuum

Different data	Different code	Different	Different
	or software	methods	context or
			country



Reproducibility

Replicability

Generalizability

- Narrow Replication (Pesaran 2003)
- Pure Replication (Hamermesh 2007)
- Verification (Clemens 2015)

- Wide Replication (Pesaran 2003)
- Statistical Replication (Hamermesh 2007)
- Reproduction/Reanalysis (Clemens 2015)

- Wider Replication (Pesaran 2003)
- Scientific Replication (Hamermesh 2007)
- Reanalysis/Robustness (Clemens 2015)

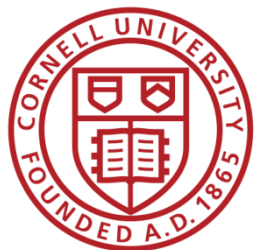
Progress



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



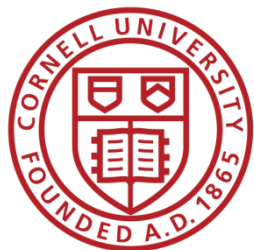
Progress

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



INSTITUT FÜR ARBEITSMARKT- UND
BERUFSFORSCHUNG
Die Forschungseinrichtung der Bundesagentur für Arbeit



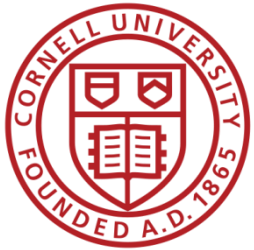


Progress

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

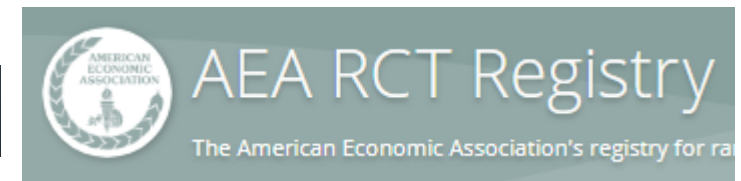
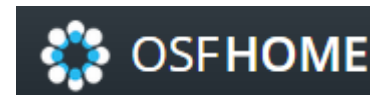
RePEc





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



More
recently...



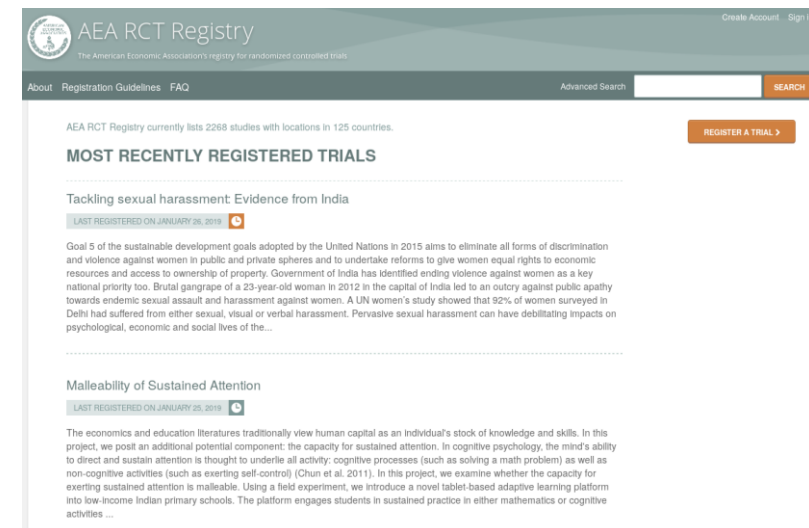
Second round (2012-)

- **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 (also EJ, Restud)



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

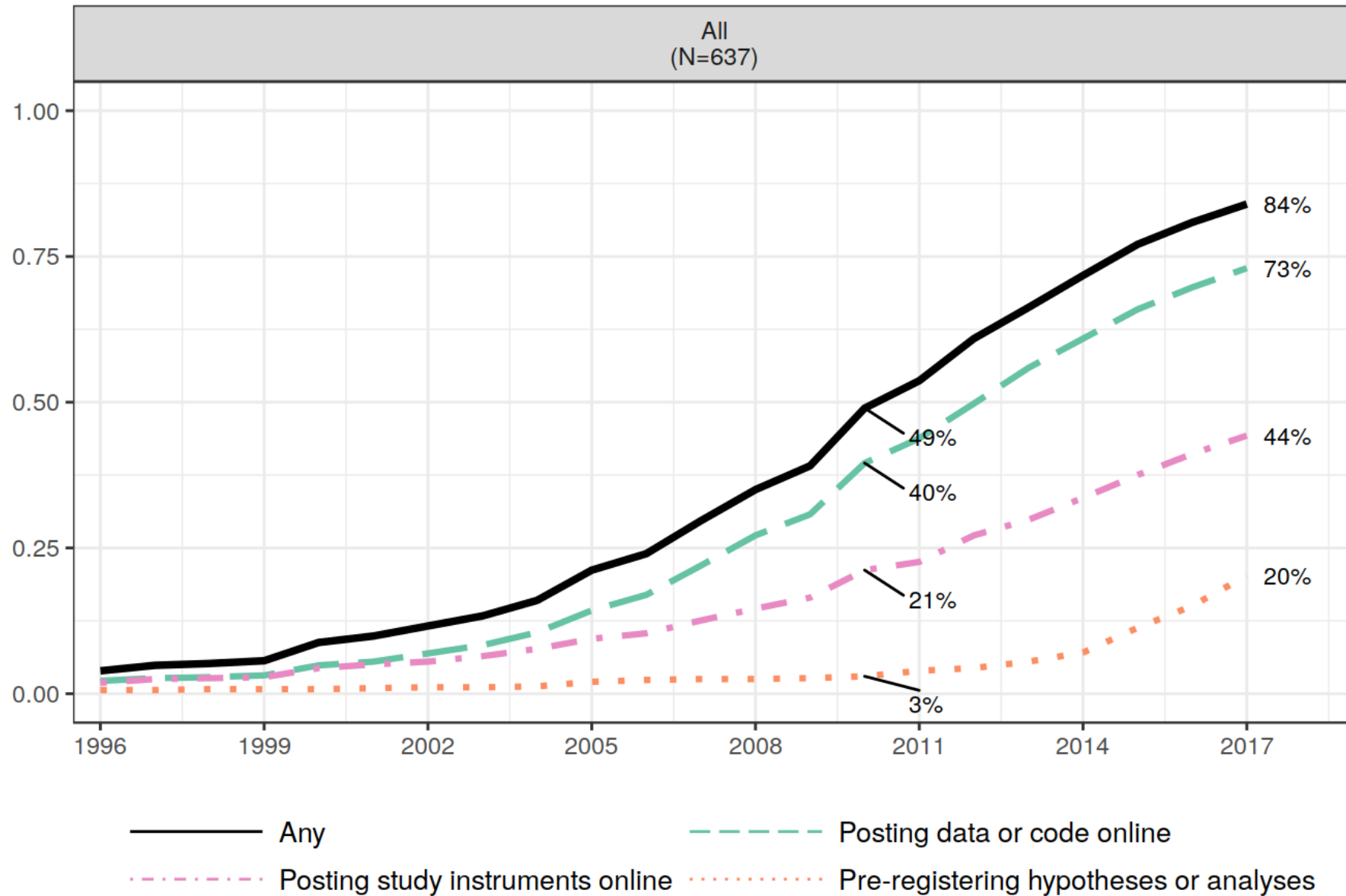
Search

[Advanced Search](#)





Share of Published Authors (PhD < 2010) Adopting Practice



Restricted-
Access Data
Pose Problems



Challenges

- **Documentation**

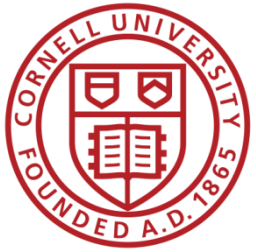
- How can others learn about the data?

- **Verifiability**

- How can others obtain access?

- **Persistence**

- How are data and programs preserved?



Challenges

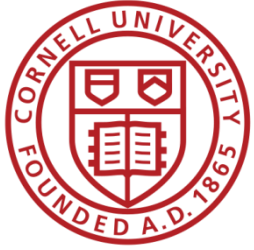
Documentation

How can others learn about the data?

- How can others obtain access?

- **Persistence**

- How are data and programs preserved?



Lack of proper dataset citation or identification



Economics makes wide use of public-use data

- **Macrodata:**

“We use data downloaded from the Bureau of Economic Analysis...”

- **Microdata:**

“... this paper uses data from the Current Population Survey...”



Relevant example

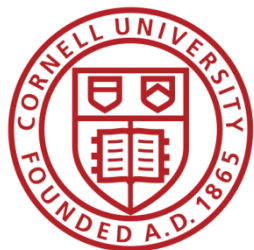
Using Business Dynamics Statistics (BDS), we follow cohorts of firms, starting from their year of entry. The data span all US nongovernment sectors and cover

Page 5 of Online Appendix:

not credible. These establishments are excluded from the change calculations in a given year” (<http://www.census.gov/ces/dataproducts/bds>). Therefore, we check whether our

The manuscript, the online appendix, and the README contain no reference to the data location or the fact that the U.S. Census Bureau publishes the Business Dynamics Statistics.

Sedláček & Sterk (2017)
10.1257/aer.20141280



This should be easy!



Problems Making RELIABLE archives

Many datasets

- Are imperfectly described
 - Very few data citations
- Are badly documented
- Have no (permanent) location defined
 - Even for data from high-profile organizations!
- All of the above



Example: BDS

Business Dynamics Statistics (BDS)

About the BDS

2016 Update!

The Business Dynamics Statistics (BDS) provides annual measures of business dynamics (such as job creation and destruction) aggregated by establishment and firm characteristics. The BDS is created from the [Longitudinal Business Database \(LBD\)](#), a [Research Data Centers](#). The use of the LBD as its source data permits tracking establishments and firms over time.

BDS data tables show key economic data:

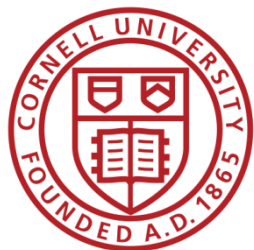
- Employment – job creation and destruction
- Job expansions and contractions
- Number of establishments
- Establishment openings and closings
- Number of startups and firm shutdowns

The BDS se

- Annu
-
-
-
- Cove

- No DOI or permanent URL
- No suggested data citation
- No versioned data releases

Doesn't make it easy!



<sidenote>



What is a data citation?

Benefits for data producers:

- provides proper attribution and credit
- creates a bibliographic "trail", connecting publications and supporting data
- demonstrates the impact of their work and establishes research data as an important contribution to the scholarly record

Benefits for data users:

- citation makes it easier to find datasets
- supports persistence of datasets
- encourages the reuse of data for new research questions

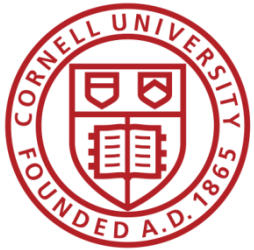


What is a data citation?

Benefits for everyone:

- increases transparency and reproducibility

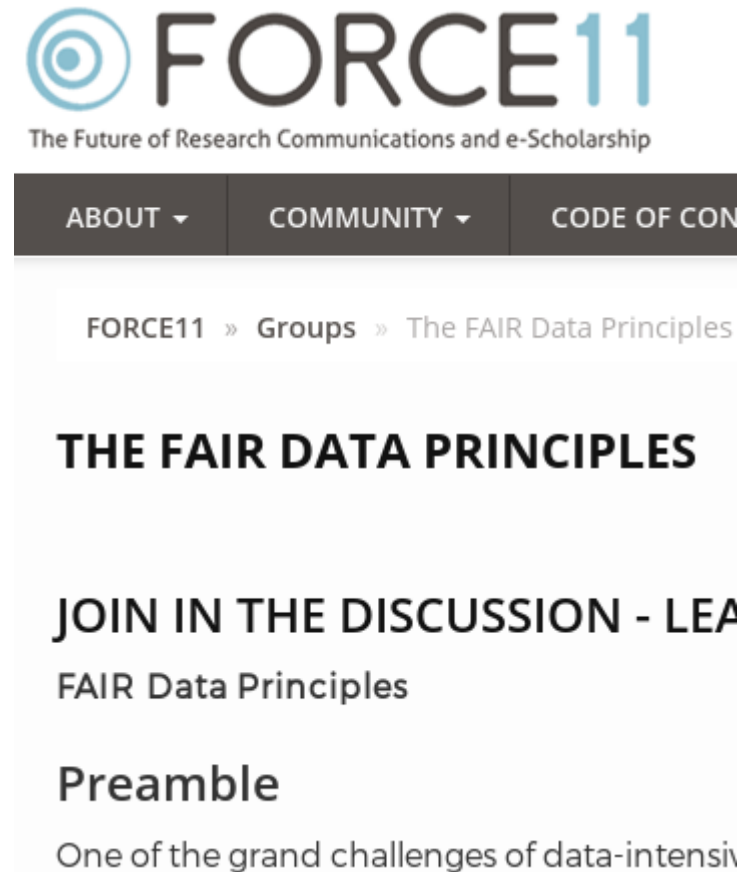
Smith, Tom W., Peter V. Marsden, and Michael Hout. 2011. *General Social Survey, 1972-2010 Cumulative File*. ICPSR31521-v1. Chicago, IL: National Opinion Research Center. Distributed by Ann Arbor, MI: Inter-university Consortium for Political and Social Research.
doi:10.3886/ICPSR31521.v1



Action: Data citations and metadata

What is **FAIR**?

- **F**indable,
- **A**ccessible,
- **I**nteroperable, and
- **R**e-usable





FAIR principles rely on metadata

Subject Terms ?

Do not copy/paste multiple terms into this field. Terms must be entered

× Rural roads

JEL Classification ?

× J43 Agricultural Labor Markets × O12 Microeconomic Analyses of

× O18 Urban, Rural, Regional, and Transportation Analysis • Housing

Manuscript Number ?

AER-2018-0268.R1 [✎ edit](#) [✕ remove](#)

Geographic Coverage ? [+ add value](#)

India [✎ edit](#) [✕ remove](#)

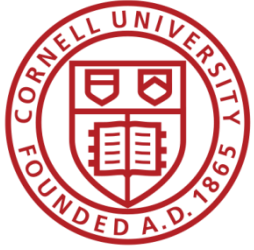
Time Period(s) ? [+ add value](#)

2000 – 2013 [✎ edit](#) [✕ remove](#)

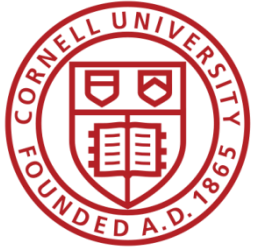
Collection Date(s) ? [+ add value](#)

Universe ?




Villages in India without paved roads in 2000. [✎ edit](#) [✕ remove](#)



</sidenote>



Example: BDS

Name	Last modified	Size	Description
 Parent Directory		-	
 2015/	14-Sep-2017 09:41	-	
 2016/	11-Oct-2018 14:19	-	
 estab/	06-Sep-2016 07:50	-	
 firm/	06-Sep-2016 07:47	-	

Note on versioned releases

- Recent releases are for “additional years of coverage” (2016 = data for 2016)
- Does not account for the possibility of re-releases or corrections of data



Especially problematic for confidential data



Example: LBD

Economic Studies (CES)

Research
opportunities

Research
Programs

Publications
and Reports

FAQs

Longitudinal Business Database (LBD)

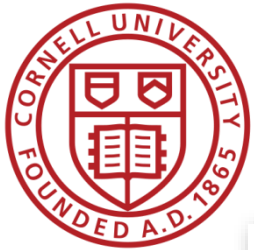
The Longitudinal Business Database (LBD) restricted-use microdata is accessible only to qualified researchers. The LBD provides important new insights about business formation and growth, the natural connections to credit markets and financing to name a few. The LBD and [Business Dynamic Statistics \(BDS\)](#) both the establishment and the firm level over a long period of time.

The LBD is a census of business establishments and firms in the U.S. with paid employees comp

LBD Restricted-Use Microdata

Sector	Economic – Establishment Surveys and Data
Industry	Economy-wide
Frequency	Annually
Sponsor	Census Bureau
Unit of Enumeration	Establishment
Availability	1976–2013
Observations	Over 8.5 million in 2013
Related Links	Business Dynamics Statistics (BDS) Synthetic Longitudinal Business Database (SynLBD)
More Information	Jarmin, Ron S. and Javier Miranda. 2002. " The Longitudinal Business Database "

- No DOI
- No suggested data citation
- No public summary statistics
- No public codebook



It can be done

6. Publishing the results

Researchers' publications based on FDZ data must include a reference to the datasets used (name of the datasets and mode of the data access).



[Citation of the data in publications](#)

Please submit one copy of every publication (including so-called 'grey literature') in digital or printed form to the FDZ.

Data (version BHP 7516)

'This study uses the weakly anonymous Establishment History Panel 1975-2016, DOI: 10.5164/IAB.BHP7516.de.en.v1. Data access was provided via on-site use at the Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB) and/or remote data access.'



Formatted citation

DOI Citation Formatter

Paste your DOI:

10.5164/IAB.BHP7516.de.en.v1



For example 10.1145/2783446.2783605

Select Formatting Style:

apa



Schmucker, A., Eberle, J., Ganzer, A., Stegmaier, J., & Umkehrer, M. (2018). Establishment History Panel 1975-2016 (Version v1) [Data set]. Forschungsdatenzentrum der Bundesagentur für Arbeit (BA) im Institut für Arbeitsmarkt- und Berufsforschung (IAB). <https://doi.org/10.5164/iab.bhp7516.de.en.v1>

Format

Schmucker, A., Eberle, J., Ganzer, A., Stegmaier, J., & Umkehrer, M. (2018). Establishment History Panel 1975-2016 (Version v1) [Data set]. Forschungsdatenzentrum der Bundesagentur für Arbeit (BA) im Institut für Arbeitsmarkt- und Berufsforschung (IAB). <https://doi.org/10.5164/iab.bhp7516.de.en.v1>

Copy to clipboard

Do you want to integrate this service? Check the [Documentation](#)



Documenting access

Just as important as describing what data you used: how can others access the data?

- May be the same way you accessed the data

- May differ, have changed

How to keep up to date?

- Link to the agency page (“Application portal”)

The screenshot shows the top portion of the United States Census Bureau's Economic Studies (CES) website. The header is dark blue with the Census Bureau logo and navigation links for TOPICS, GEOGRAPHY, LIBRARY, and DATA. Below the header is a breadcrumb trail: > Industry > Center for Economic Studies > RDC Research Opportunities > How to Apply. The main heading is "Economic Studies (CES)" in large blue font. Below this is a green navigation bar with four buttons: "RDC Research Opportunities", "Research Programs", "Publications and Reports", and "FAQs". The "How to Apply" section is highlighted. It contains a paragraph stating that CES and RDCs consider proposals from qualified researchers in social science disciplines. It lists five criteria for proposed projects: provide benefit to Census Bureau programs, demonstrate scientific merit, require non-public data, be feasible given the data, and pose no risk of disclosure. The "Developing and Submitting a Research Proposal" section is also visible, starting with the instruction to consult the RDC administrator at the location where the project will be based. It lists two documents for download: "Research Proposal Guidelines" (602 KB) and "Proposal Registration Form [RTF]" (137 KB).

United States
Census
Bureau

TOPICS
Population, Economy

GEOGRAPHY
Maps, Products

LIBRARY
Infographics, Publications

DATA
Tools, Development

> Industry > Center for Economic Studies > RDC Research Opportunities > How to Apply

Economic Studies (CES)

RDC Research Opportunities | Research Programs | Publications and Reports | FAQs

How to Apply


CES and the RDCs consider proposals from qualified researchers in social science disciplines consistent with the subject of the data any time. All researcher access to restricted-use data occurs at secure [Federal Statistical Research Data Centers](#).

Proposed projects must

- Provide benefit to Census Bureau programs
- Demonstrate scientific merit
- Require non-public data
- Be feasible given the data
- Pose no risk of disclosure

Developing and Submitting a Research Proposal

Before preparing your research proposal, consult

- RDC administrator at the [location](#) where you want to base your project regarding access fees and the content of the data
- [Research Proposal Guidelines](#)  (602 KB)
 - [Proposal Registration Form \[RTF\]](#) (137 KB)



Documenting access

“The Longitudinal Business Database can be accessed by authorized users through the Federal Statistical Research Data Centers. For more information, see <https://www.census.gov/ces/rdcresearch/howtoapply.html>

United States[™]
Census
Bureau

TOPICS
Population, Economy

GEOGRAPHY
Maps, Products

LIBRARY
Infographics, Publications

DATA
Tools, Develop

[Industry](#) > [Center for Economic Studies](#) > [RDC Research Opportunities](#) > [How to Apply](#)

Economic Studies (CES)

RDC Research Opportunities

Research Programs

Publications and Reports

FAQs

How to Apply


CES and the RDCs consider proposals from qualified researchers in social science disciplines consistent with the subject any time. All researcher access to restricted-use data occurs at secure [Federal Statistical Research Data Centers](#).

Proposed projects must

- Provide benefit to Census Bureau programs
- Demonstrate scientific merit
- Require non-public data
- Be feasible given the data
- Pose no risk of disclosure

Developing and Submitting a Research Proposal

Before preparing your research proposal, consult

- RDC administrator at the [location](#) where you want to base your project regarding access fees and the content
- [Research Proposal Guidelines](#)  (602 KB)
 - [Proposal Registration Form](#) [RTF] (137 KB)



Documentating access

“Data can be accessed via on-site use at the Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB. For more information and requirements, see https://fdz.iab.de/en/FDZ_Data_Access/FDZ_On-Site_Use.aspx



RESEARCH DATA CENTRE (FDZ)
of the German Federal Employment Agency (BA)
at the Institute for Employment Research (IAB)

[Home](#) | [Newsletter](#) | [Jobs](#) | [Contact](#) | [Data Privacy](#) | [Imprint](#)

About us
Overview of Data
Establishment Data
Individual Data / Household Data
Integrated Establishment and Individual Data
External data
Data Archive
Data Access
On-Site Use
Locations
Remote Data Access /

Data Access: On-site Use

The use of weakly anonymous data is subject to restrictions concerning [data](#); these regulations the data can be analyzed only via on-site use. For this purpose separate workplaces for guest Researchers at different [locations](#).

The FDZ offers advisory service by its staff only at the location in Nuremberg.

- [How to obtain data access to the FDZ data via On-site Use?](#)
- [How to modify an existing FDZ agreement?](#)
- [Data use for students](#)
- [Specifics on the data access at FDZ locations in USA, Canada and UK](#)



How to improve?

- Agencies should provide template language for researchers to use
- Agencies should provide data citations (see IAB example)
- Agencies should provide DOI
- Agencies should provide public documentation
 - Database schema/ structure/ codebook
 - Generic summary statistics



Challenges

- Documentation

Verifiability

How can others obtain access?

- Persistence

- How are data and programs preserved?



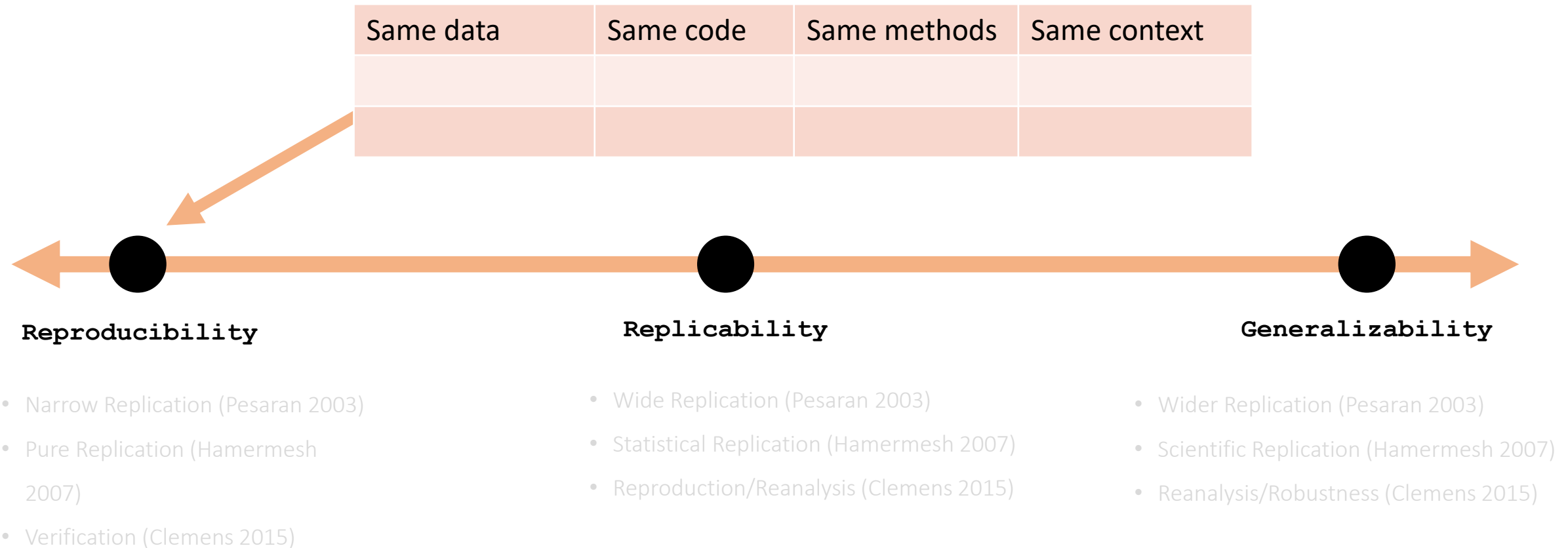
Reproducibility problems

1. Is it reproducible in the first place?
2. How can I prove it to others?

Not enough
articles are
reproducible

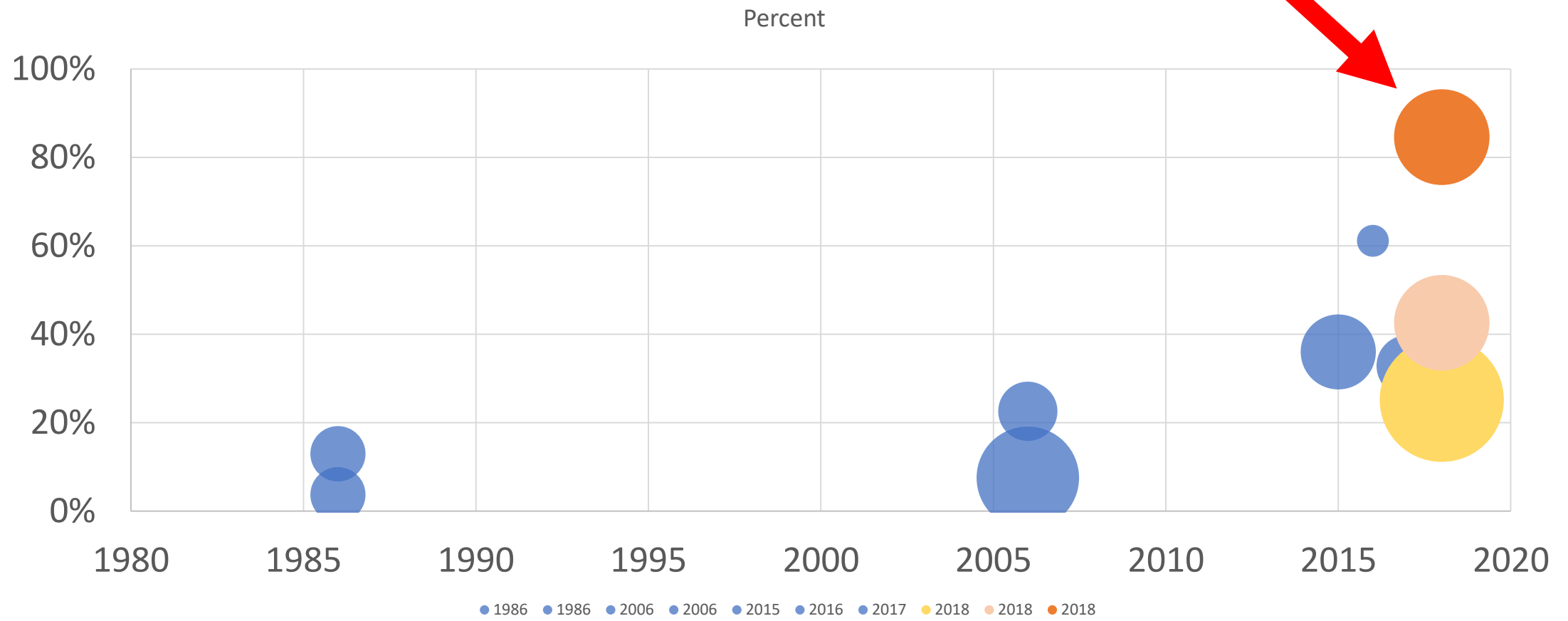


Replication continuum





Results?



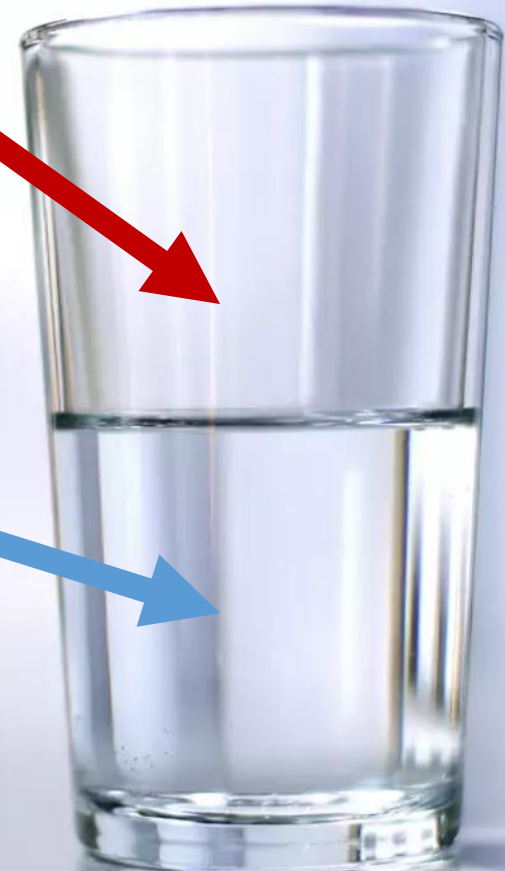


In a nutshell

- **40%** use restricted-access data
- **25%** use public-use data and are mostly or completely reproducible
- **25%** use public-use data and are only partially reproducible
- **10%** fail to yield useful results

It's only ½ full!

Hey, it's not empty!





Reproducibility is harder than it should be

- Often done piecemeal
 - At different times
 - By different people
- Software versions
 - Stata 9? 15? 42?
 - rdrobust 2014? 2016? 2018 bug fix?
- Compilers and exotic software



Most frequent problem

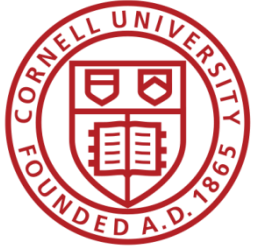


Making USEFUL archives

- From analysis of code from 1996 to 2003 (MMH2006):

“Other authors seem to think that the entire world shares the exact same hard drive layout, with “C:\MYDATA\MYPROJECT\” **sprinkled liberally** throughout their code. Of course, a would-be replicator has to **find and change all these**.”

“The author might not realize all the data/subroutine files that his code utilizes, and **forget to include** said data/subroutine in his replication files.”



Especially true in FSRDC...



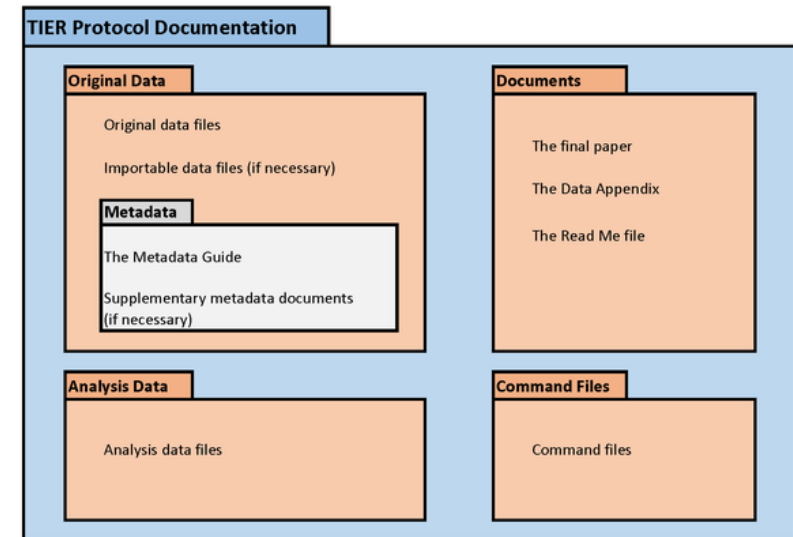
Issues in RDCs

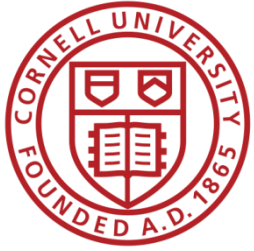
- Changes in file locations and directory
- Changes in computing architecture
- Not specific to RDCs:
 - Change in RA team
 - Data flow management
 - Broken dependencies



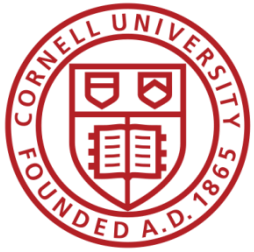
Use best practices for reproducibility

- TIER Protocol
- Versioning of code (and data!)
- Robust programming techniques (configuration files, modular code)
- Secure programming techniques! (no confidential parameters in core code)

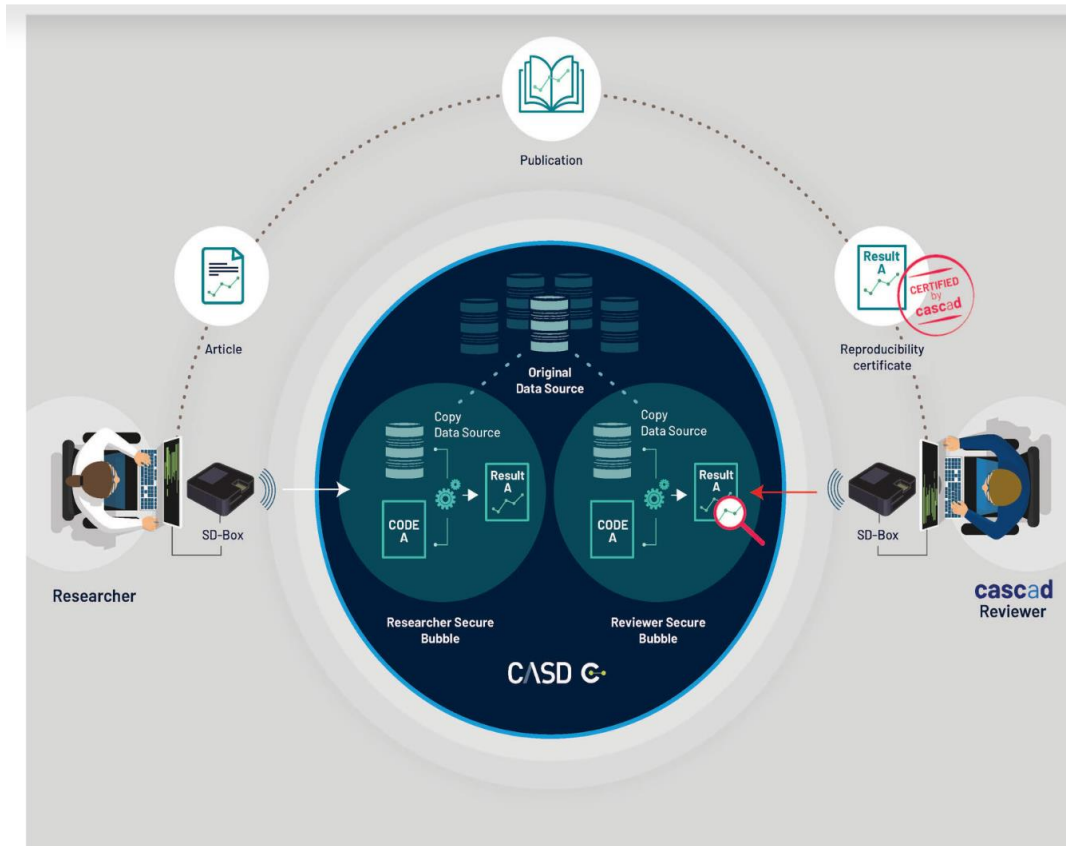




How can others verify reproducibility?



How can others verify reproducibility?



 **cascader**
*the first certification
agency for scientific
code & data*

A cascader certification allows researchers to signal the reproducibility nature of their research to their peers



Verification service

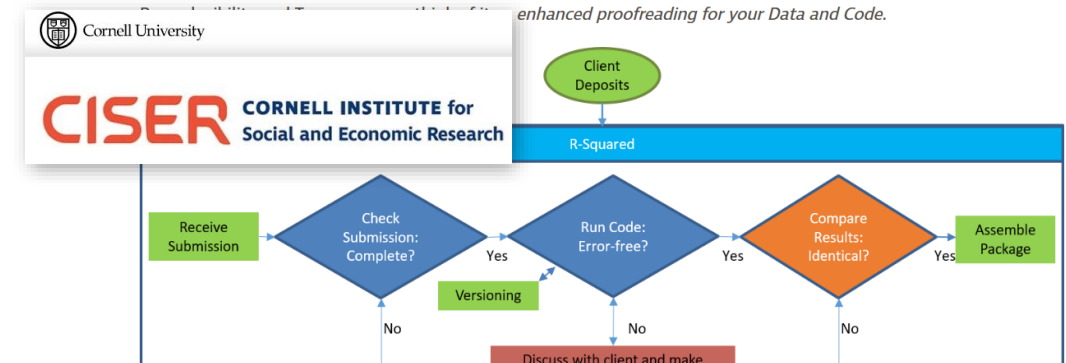
- Permanently accredited staff with access to confidential data
- On-demand, arms-length verification of code and results
- Effectiveness rests on credibility of the service
- Similar to existing services that perform pre-submission verification within universities and research institutes, but with access to confidential data

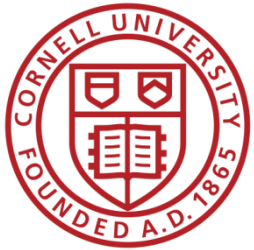


Home > Research > **Results Reproduction (R-squared)**

RESULTS REPRODUCTION (R-SQUARED)

Results Reproduction (R-Squared) is a service that computationally reproduces the results of your research to ensure enhanced proofreading for your Data and Code.





Embed within disclosure avoidance system



FORSCHUNGSDATENZENTRUM
der Bundesagentur für Arbeit im Institut für
Arbeitsmarkt- und Berufsforschung

[Start](#) | [Newsletter](#) | [Stellenangebote](#)

[Wir über uns](#)

Das Forschungsdatenzei

- Researchers develop code interactively (thin client, web application)
- Request to have results released
- IAB, not researcher, re-executes codes, and releases results produced by the code
- **Implicitly certifies reproducibility!**



Ad-hoc or peer-to-peer

- Journal Data Editor may ask for suggestions of reviewers with access to the same data
- Journal Data Editor may ask for permission for reviewers to access the data
- Protocol to be followed
- Case by case
- Effort?

AEA Data and Code Guidance



**AMERICAN
ECONOMIC
ASSOCIATION**

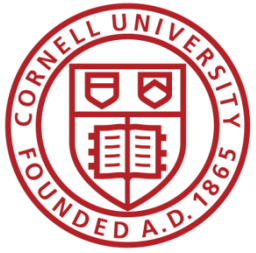
Guidance for authors wishing to create data and code supplements, and for replicators.

Protocol for Third-party Verifications

This protocol describes how third parties can, at the request of the AEA Data Editor, conduct a reproducibility check of materials that are part of an AEA publication.

Alternate protocols are possible, but should be verified with the AEA Data Editor prior to engaging any resources.

Preliminaries



Other options

- Cryptographic signatures
- Blockchain
- All of the previous



Image by [VIN JD](#) on Pixabay



This is an unsolved problem.



Challenges

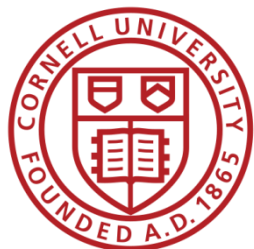
- **Documentation**

- How can others learn about the data?

- **Verifiability**

Persistence

How are data and programs preserved?



A public-use data solution

J Econom. Author manuscript; available in PMC 2012 Mar 1.

Published in final edited form as:

J Econom. 2011 Mar 1; 161(1): 82–99.

doi: [10.1016/j.jeconom.2010.09.008](https://doi.org/10.1016/j.jeconom.2010.09.008)

PMCID: PMC3079891

NIHMSID: NIHMS246950

National Estimates of Gross Employment and Job Flows from the Quarterly Workforce Indicators with Demographic and Industry Detail

[John M. Abowd](#) and [Lars Vilhuber](#)

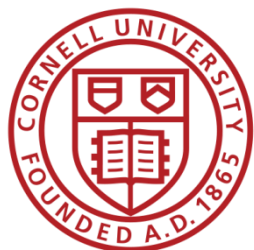
[Author information](#) ► [Copyright and License information](#) ►

Abstract

[Go to:](#) ☒

The Quarterly Workforce Indicators (QWI) are local labor market data produced and released every quarter by

No confidential data were used in this paper. All public-use Quarterly Workforce Indicators data can be accessed from <http://www.vrdc.cornell.edu/news/data/qwi-public-use-data/>. The national indicators developed in this paper can be accessed from <http://www.vrdc.cornell.edu/news/data/qwi-national-data/>. We are grateful for the comments and suggestions of many of our colleagues, past and present, too numerous to list here and thus listed at the website above and in the working paper version of this article. The opinions expressed in this paper are those of the authors and not the U.S. Census Bureau nor any of the research sponsors.



An example: not cited...

J Econom. Author manuscript; available in PMC 2012 Mar 1.

Published in final edited form as:

J Econom. 2011 Mar 1; 161(1): 82–99.

doi: [10.1016/j.jeconom.2010.09.008](https://doi.org/10.1016/j.jeconom.2010.09.008)

PMCID: PMC3079891

NIHMSID: NIHMS246950

National Estimates of Gross Employment and Job Flows from the Quarterly Workforce Indicators with Demographic and Industry Detail

[John M. Abowd](#) and [Lars Vilhuber](#)

[Author information](#) ► [Copyright and License information](#) ►

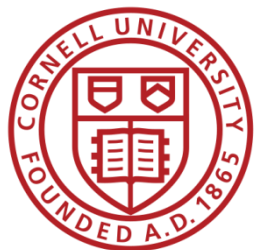
Abstract

[Go to:](#) ☒


The Quarterly Workforce Indicators (QWI) are local labor market data produced and released every quarter by

Press for the NBER; 2009. pp. 149–230.

5. Abowd JM, Vilhuber L. The sensitivity of economic statistics to coding errors in personal identifiers. Journal of Business and Economic Statistics. 2005;23(2):133–152.
6. Abowd JM, Zellner A. Estimating Gross Labor Force Flows. Journal of Business and Economic Statistics. 1985;3:254–283.



We went back, archived it



 **Dataverse**

Search About

Lars Vilhuber Dataverse (Cornell University)

Harvard Dataverse > Lars Vilhuber Dataverse >

Replication data for: National estimates of gross employment and job flows from the Quarterly Workforce Indicators with der

 Metrics  4 Downloads

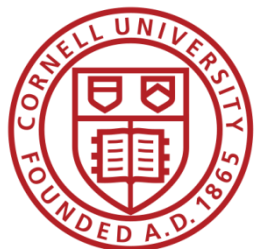
Replication data for: National estimates of gross employment and job flows from the Q Indicators with demographic and industry detail

Abowd, John M.; Vilhuber, Lars, 2014, "Replication data for: National estimates of gross employment and job flows from the Quarterly Workforce Indicators with demographic and industry detail", <http://dx.doi.org/10.7910/DVN/27923>, Harvard Dataverse, V2

If you use these data, please add this citation to your scholarly resources. Learn about [Data Citation Standards](#).

Description

The Quarterly Workforce Indicators are local labor market data produced and released by the Bureau of Economic Analysis. Unlike any other local labor market series produced in the U.S. or the rest of the world, the QWIs provide flows for workers (accession and separations), jobs (creations and destructions) and earnings (by sex), economic industry (NAICS industry groups), and detailed geography (county, Census Workforce Investment Area, as well as experimental, unreleased block-level estimates). The QWIs are the first national compilation of the existing public-use data (and only those public-use data) to construct the first national important enhancement to existing series because they include demographic and industry information compiled from data that have been integrated at the micro-level by the Longitudinal Employment Dynamics (LEDS) project.



We went back, archived it, linked it back

 Dataverse

 About

Keyword

Employment Dynamics

Topic Classification

Economics

Related Publication

John M. Abowd and Lars Vilhuber, "National estimates of gross employment and job flows from the Quarterly Work with demographic and industry detail," Journal of Econometrics, vol. 161, iss. 1, pp. 82-99, 2011. doi:

10.1016/j.jeconom.2010.09.008 <http://www2.vrdc.cornell.edu/news/data/qwi-national-data/>

John M. Abowd and Lars Vilhuber, "National estimates of gross employment and job flows from the Quarterly Work with demographic and industry detail," Journal of Econometrics, vol. 161, iss. 1, pp. 82-99, 2011. doi:

10.1016/j.jeconom.2010.09.008 <http://www2.vrdc.cornell.edu/news/data/qwi-national-data/>

John M. Abowd and Lars Vilhuber, "National estimates of gross employment and job flows from the Quarterly Work with demographic and industry detail (with color graphs)," Center for Economic Studies, U.S. Census Bureau, Wc 11, 2010. <http://ideas.repec.org/p/cen/wpaper/10-11.html>

Producer

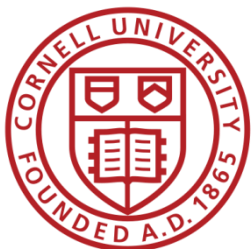
Labor Dynamics Institute (Cornell University) (LDI) <http://www2.vrdc.cornell.edu/news/data/qwi-national-data/>





But: Encourage Best Practices

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



Deposit as soon as you can

OPEN ICPSR

Find Data

Share Data

openICPSR Repositories ▾

[Find Data](#) / [Replication data: Total Error and Variability Measures for QWI and LODES](#)

Replication data: Total Error and Variability Measures

Principal Investigator(s): ⓘ Kevin L. McKinney, United States Department of Commerce. Bureau of the C
Cornell University; John M. Abowd, Cornell University; John M. Abowd, United States Department of Comm

Version: ⓘ V1

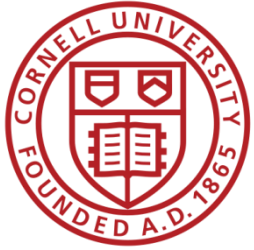
Version Title: ⓘ Original version

Name ⓘ	File Type ⓘ
 README.txt	text/x-web-m

Project Citation:

McKinney, Kevin L., Green, Andrew S., Vilhuber, Lars, Abowd, John M., and Abowd, John M. Replication data: Total Error and Variability Measures for QWI and LODES. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2017-12-15. <https://doi.org/10.3886/E100590V1>

Persistent URL: <http://doi.org/10.3886/E100590V1>



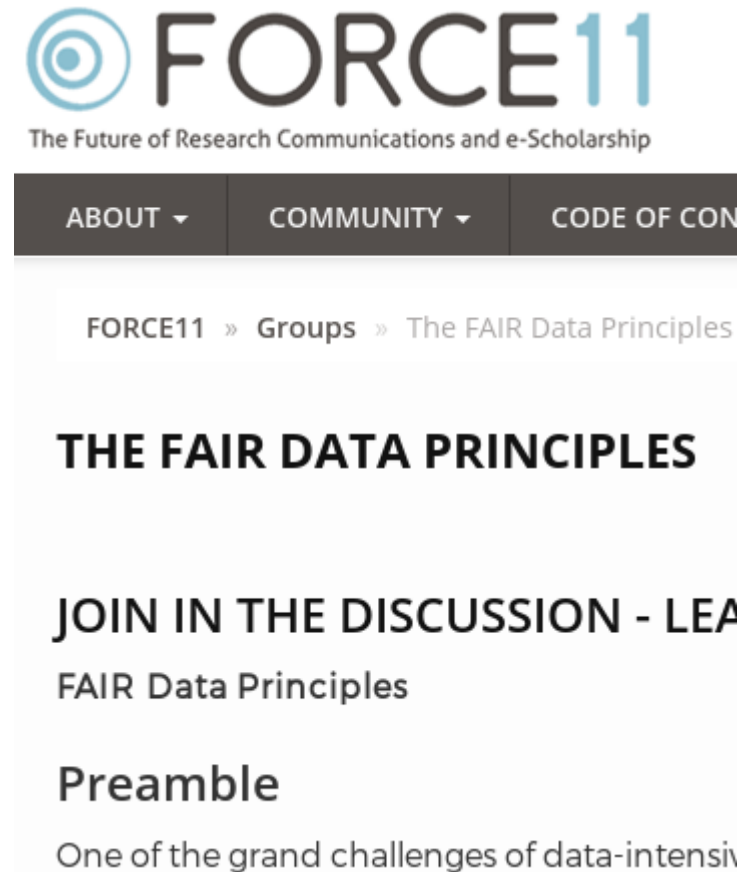
This doesn't work in RDC!



Action: Data citations and metadata

What is **FAIR**?

- **F**indable,
- **A**ccessible,
- **I**nteroperable, and
- **R**e-usable





Preserving data in restricted access centers

- Use existing procedures
- Release as much non-sensitive information to public repositories
- Reveal the existence of confidential/sensitive information



Use existing procedures

- Version your code (**git or svn work without Gitlab or Github!**)
- Leverage disclosure avoidance procedures
 - tie “code releases” and “tables” to specific “disclosure avoidance numbers”
- Allow (10-year) backup to “archive” files
 - Make those files findable!
 - Tie to some DOI or findable tag



Some guidance



Data and Code Guidance by Data Editors

Guidance for authors wishing to create data and code supplements, and for replicators.

UNF:6:Upe25NYAZwR+6VsDt5X2IQ==

Challenges in Hosting of Data and Code at Restricted-Access Data Centers

Users of restricted-access data centers (RADC, such as [FSRDCs](#), [CASD](#), etc.) face certain challenges in the handling of data and code as described in this document:

- researchers (end-users) may not be able to provide DOI or similar persistent identifiers for some data
- researchers may not be able to discern the preservation policy for certain data sets
- researchers may not be able to remove all code from the center, or such removal is subject to restrictions
- data citation guidance may be lacking, or may not be obvious (see [Data Citation Guidance](#) for general guidance)

A few guidelines

- [https://social-science-data-editors.github.io/guidance/Requested information hosting.html#challenges-in-hosting-of-data-and-code-at-restricted-access-data-centers](https://social-science-data-editors.github.io/guidance/Requested%20information%20hosting.html#challenges-in-hosting-of-data-and-code-at-restricted-access-data-centers)



Some more extensive guidance

Making Confidential Data Part of Reproducible Research

Lars Vilhuber, *Cornell University*
Carl Lagoze, *University of Michigan*

<https://digitalcommons.ilr.cornell.edu/ldi/41/>



Cornell University
ILR School

Cornell University ILR School
DigitalCommons@ILR

Labor Dynamics Institute

Centers, Institutes, Programs

8-21-2017

Making Confidential Data Part of Reproducible Research

Lars Vilhuber
Cornell University, lv39@cornell.edu

Carl Lagoze
University of Michigan, clagoze@umich.edu



Surfacing additional information

- For documents:

CES Technical Notes Series

“CES Technical Notes may contain confidential data and, thereby, disclosure is prohibited.

Researchers on approved with the correct permissions can request full text notes from (EMAIL)”

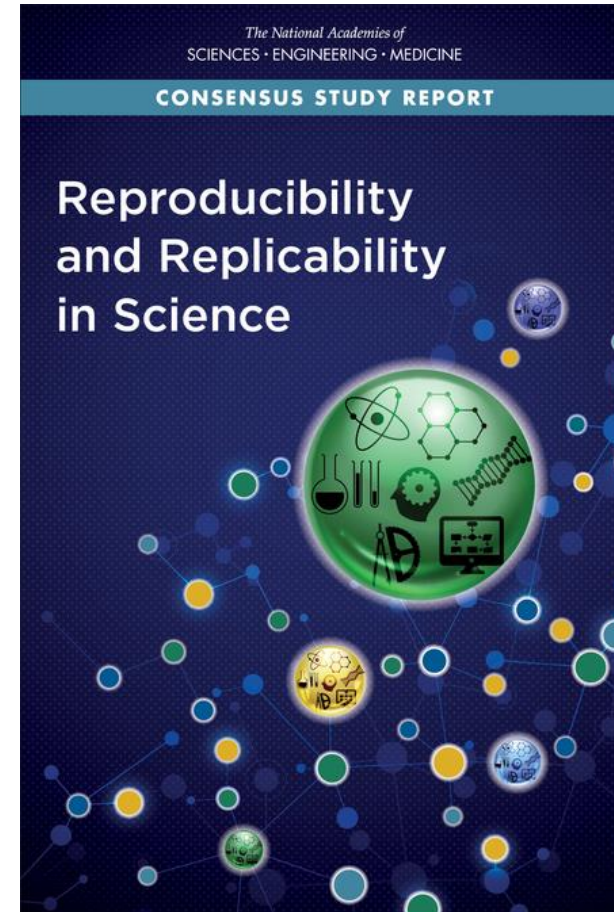
- Exploring ability to attach data and code archives

A screenshot of the IDEAS (International Development Economics and Associates) website. The header includes the IDEAS logo and navigation links for "Economic literature", "Authors", "Institutions", and "Rank". Below the header are social media icons for Twitter, Facebook, Google+, YouTube, LinkedIn, and a printer icon. The main content area displays "Center for Economic Studies, U.S. Census Bureau" and "CES Technical Notes Series". A search bar is present with the placeholder text "Search within this serial". Below the search bar are four tabs: "Publisher Info", "Serial Info", "Content" (which is selected and highlighted in dark blue), and "Corrections". Under the "Content" tab, the year "2019" is listed, followed by two entries: "19-02 Augmenting the LBD with Firm-Level Revenue" by John Haltiwanger & Ron Jarmin & Robert Kulick & Javier Miranda & Veronika Pencia, and "19-01 Duplicate Records and Cross-Year Identifier Inconsistency in Census Year S" by Lucas Threinen. Below these, the year "2018" is listed, followed by the entry "18-03 The Creation and Use of the SIPP Synthetic Beta v7.0".



Even better: repositories

- report's Recommendation 6-5 encourages NSF to create “**code and data repositories for long-term archiving and preservation of digital artifacts that support claims made in the scholarly record.**”
- These should also be implemented within non-public areas (e.g., FSRDC, etc.)



<https://doi.org/10.17226/25303>



Evolving Journal and Data Infrastructure

**Goal: Use any
repository!**
(subject to conditions)



Evolving Journal and Data Infrastructure

Treat all archives
symmetrically!



Data (and Code) Availability Statements

- A statement about where data supporting the results reported in a published article can be found
 - including unique identifiers linking to publicly archived datasets analyzed or generated during the study.
- DASs can increase transparency by providing a reason why data cannot be made (immediately) available
 - need for registration, ethical or legal restrictions, or because of an embargo period

Challenges?

Some
positives



Some random notes

- **Pre-release verification** conducted within the disclosure avoidance mechanism?
- Analogy between **grant** or **RDC proposal** and **pre-registration**
- Incentives of stats agencies: **transparency** = **credibility**
- From **pre-acceptance verification** to **pre-submission verification** (university or institute services)



Registered Reports

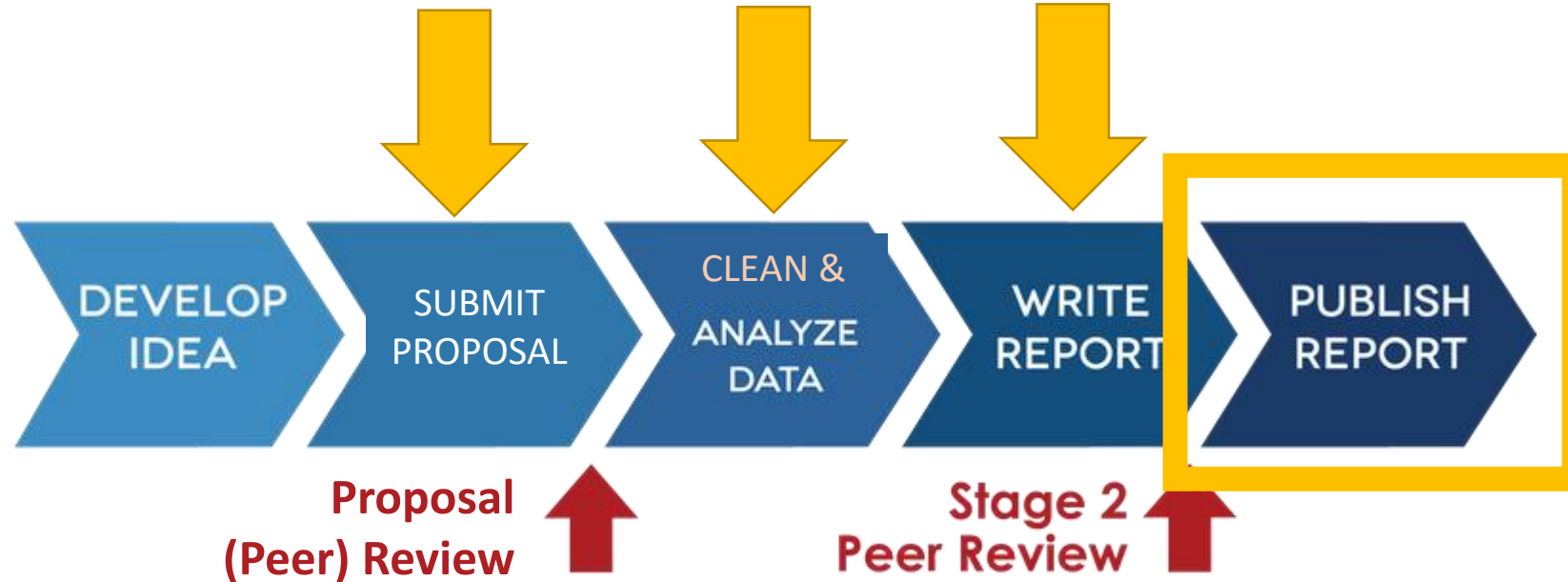
- Close cousin: Results-blind review





Registered Reports

- Close cousin: **RDC access mechanism**





Interesting thought

Could the RDC proposal and access mechanism be combined with a registered-report mechanism?



Pre-release verification conducted within the disclosure avoidance mechanism?

RDC users already submit to a hands-off (objective) code execution system (mostly) [qsub / PBS Pro]

- Learning curve
- Lots of keyboard exercise until it works



Pre-release verification conducted within the disclosure avoidance mechanism?

RDC users already submit to a hands-off (objective) code execution system (mostly) [qsub / PBS Pro]

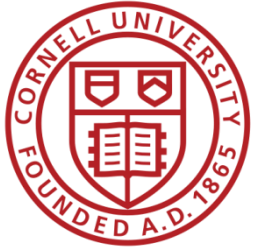
- Could code runs and their output be digitally signed?
- ... and be automatically submitted to the disclosure avoidance mechanism?

Collaboration



Goal: Transportability

Any standards, tools, methods: must be transportable across journals (no custom solutions)



Social science “guild”



Data and Code Guidance by Data Editors

Guidance for authors wishing to create data and code supplements, and for replicators.

Authors: Lars Vilhuber

This project is maintained by social-science-data-editors

Disclaimer

Unofficial guidance on various topics by Social Science Data Editors

Guidance on creating replicable data and program archives

This guidance is for the author wanting to create a replication archive.

See [Requested information](#) for the information the Data Editor may request from you, prior to the acceptance of your paper for publication.

Guidance on testing replicability of code

This guidance has two audiences:

- the author wanting to verify whether her code passes muster as a replicable archive
- the replicator wanting to verify the replicability of such an archive

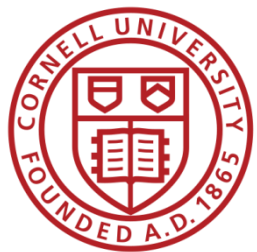
See [Verification guidance](#)

FAQ

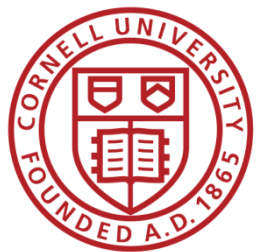
See our growing [FAQ](#). If you have questions or answers to add, please notify us by creating a [new issue](#).

[https://](https://social-science-data-editors.github.io/guidance/)
[social-science](https://social-science-data-editors.github.io/guidance/)
[-data-editors.](https://social-science-data-editors.github.io/guidance/)
[github.io/](https://social-science-data-editors.github.io/guidance/)
[guidance/](https://social-science-data-editors.github.io/guidance/)

Challenges?



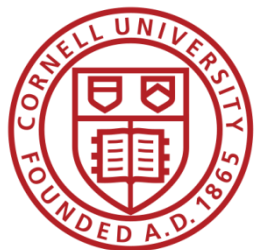
You...



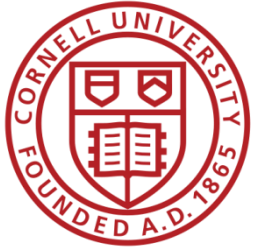
Me...



Change ingrained habits...



New skills to learn...

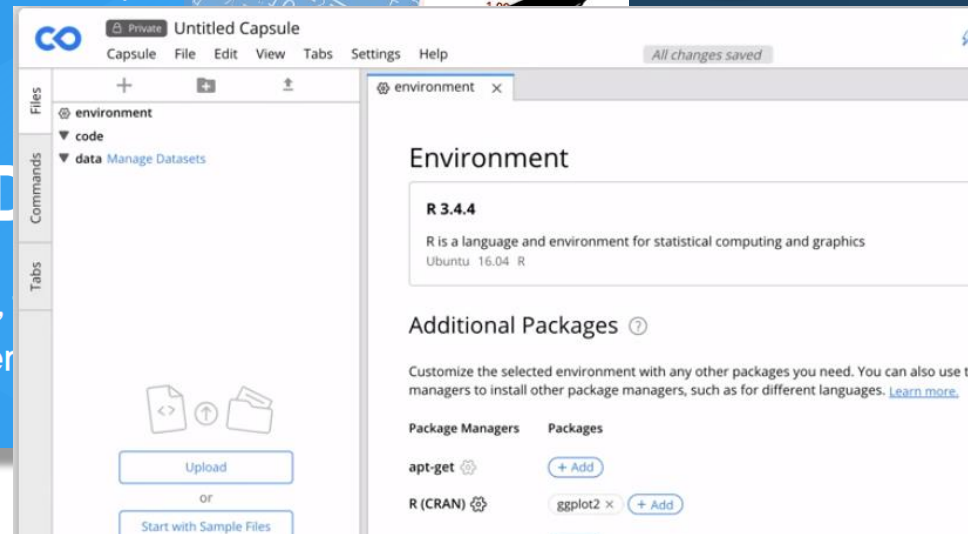


Shape Your D

Join us at DockerCon 2019,
things Kubernetes, microser

R Markdown

from R Studio



(beta)

Turn a Git repo into a collection of
notebooks

Welcome to RStudio

Do, share, teach and learn da



New methods to use ...

Delete

Univer
de M

New upload

Instructions: (i) Upload minimum one file or fill-i

Files

Recherche

MON COMPTE

Ouvrir une session

Nouvel utilisateur?

Répertoires Facultés Bibliotl

stitutionnel

```
import requests
```

```
>>> r = requests.get("https://zenodo.org/api/deposit/depositions")
>>> r.status_code
401
>>> r.json()
```

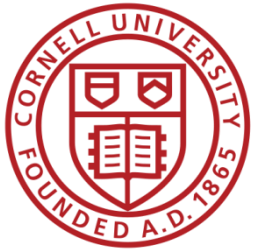
```
{
  "message": "The server could not verify that you are authorized to
```



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



Challenges

- **Documentation**

- How can others learn about the data?

- **Verifiability**

- How can others obtain access?

- **Persistence**

- How are data and programs preserved?



Possible solutions

- **Documentation**

- Better support, more consistency, new publication tools

- **Verifiability**

- Intriguing possibilities, not yet solved

- **Persistence**

- Acceptable workarounds, need for robust infrastructure

Thank you!

DOI: [10.5281/zenodo.2573123](https://doi.org/10.5281/zenodo.2573123)