



<https://forms.gle/zrY8dB8puTtud2Rt7>

<https://lars.vilhuber.com/s>

(and choose WEAI 2019)





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

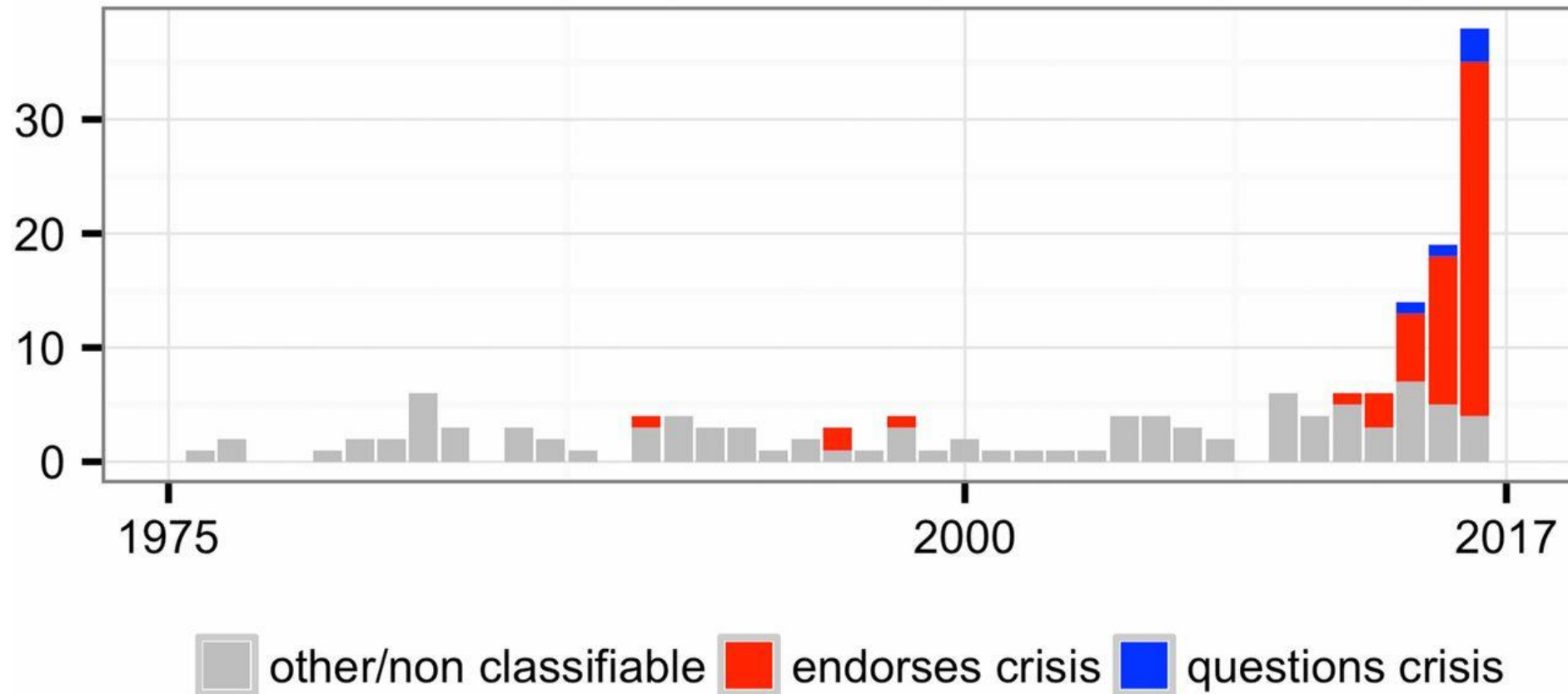
Lars Vilhuber  
Cornell University

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

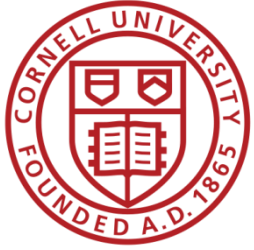
Misuse of statistics

Source: Nosek  
Sackler talk 2017



# Efficiency of scholarly discourse?

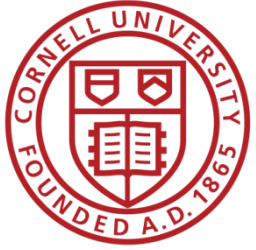
- Early publications (20<sup>th</sup> century) contained **tables of data**, and the **math** was simple (maybe)
  - **Data** became electronic, was no longer **included** or **cited**
  - **Math** was transcribed to **code**, and was no longer **included**



# Efficiency of scholarly discourse!

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

# Replication?



# Replication continuum



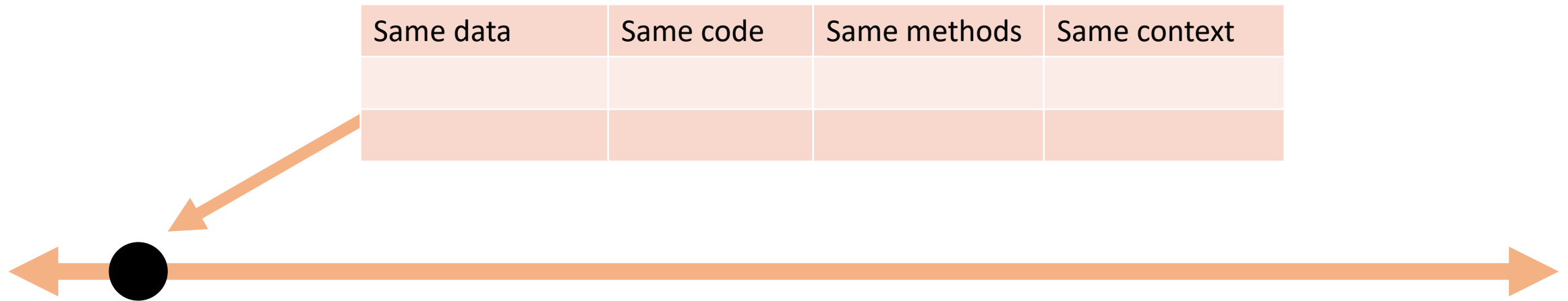
## **Reproducibility**

- 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)



# 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	<b>Different code or software</b>	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)
- Production/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)

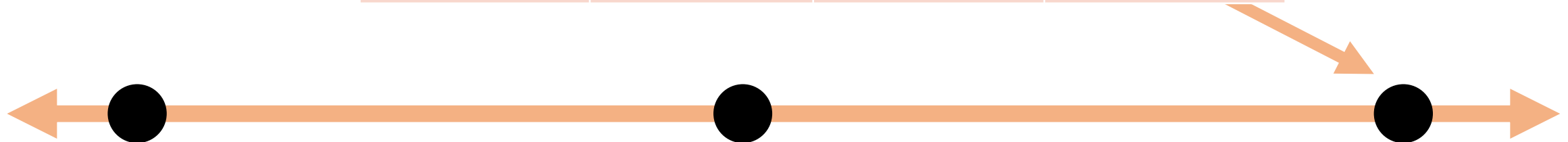
## **Generalizability**

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



# Replication continuum

<b>Different data</b>	Different code	<b>Different</b>	<b>Different</b>
	or software	<b>methods</b>	<b>context or</b>
			<b>country</b>



**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

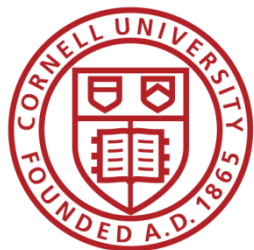


# Progress

- 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](#) 

*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*



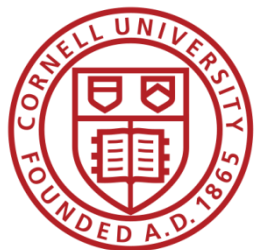
# Progress

- 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

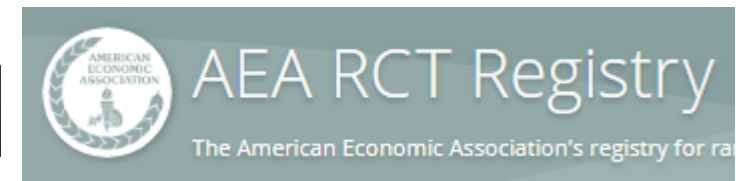
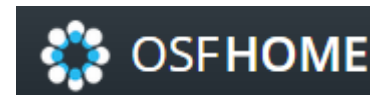
*RePEc*





# 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...



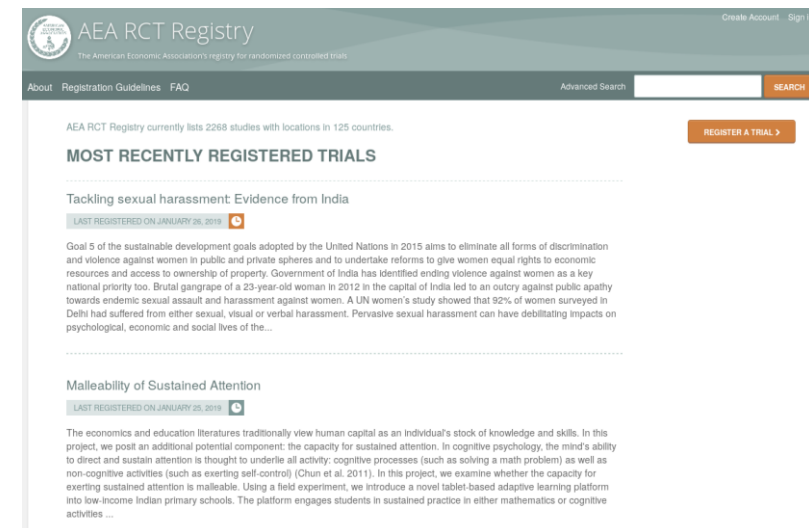
## Second round (2012-)

- **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

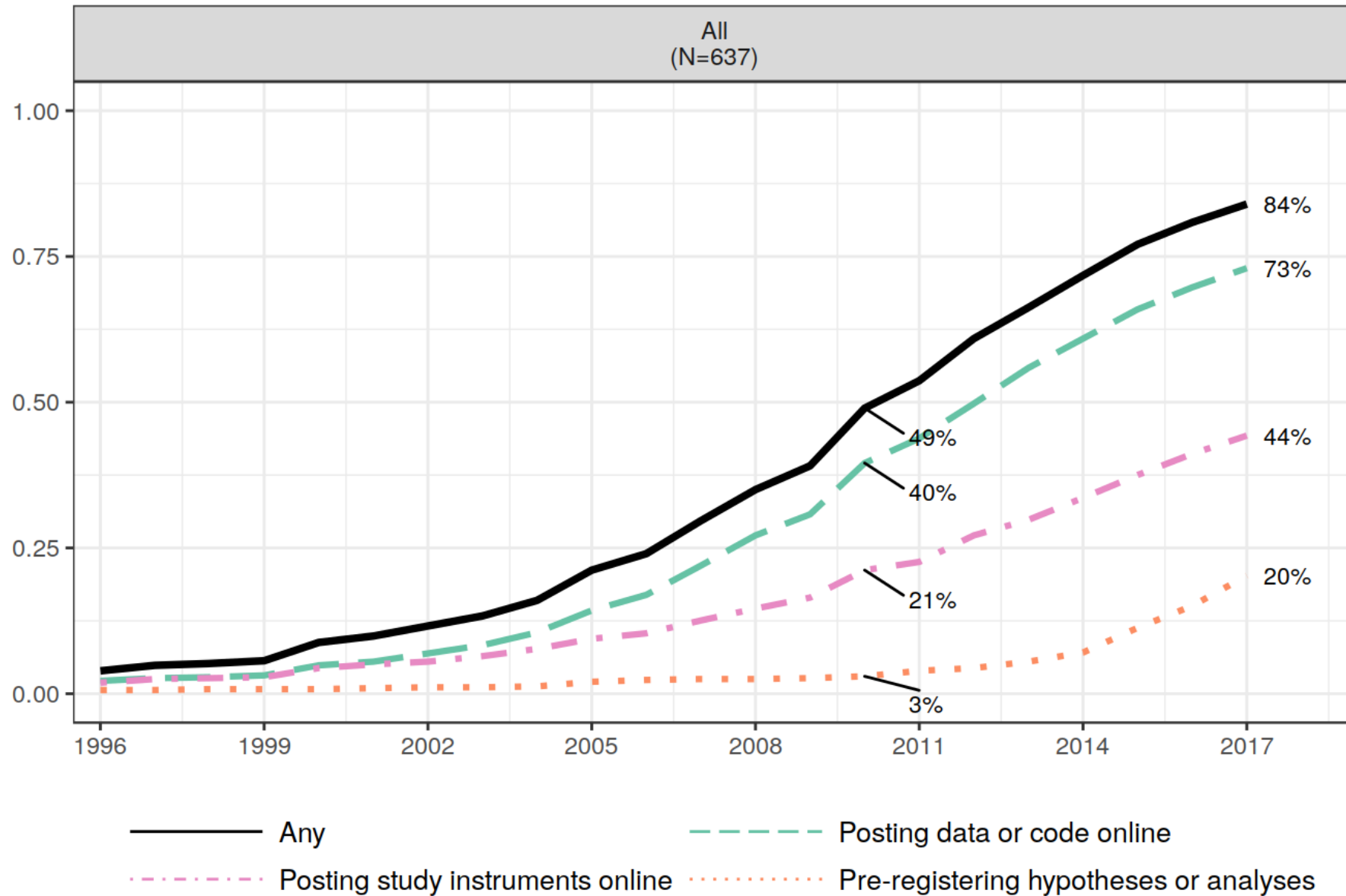
Search

Advanced Search





## Share of Published Authors (PhD < 2010) Adopting Practice



# Issues



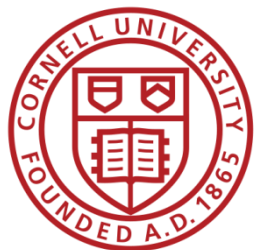
# Economics makes wide use of public-use data

- **Macrodata:**

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

- **Microdata:**

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



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

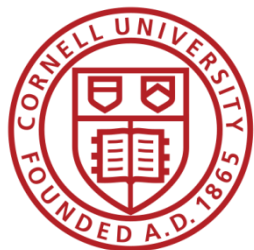


# Making USEFUL archives

- 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.”

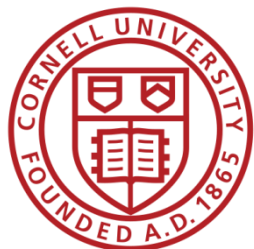


Still true today...





Let's try and do better...



# An example

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](#)

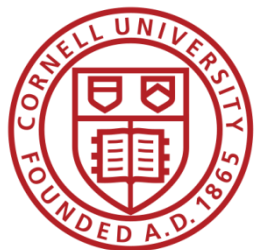
[Author information](#) ► [Copyright and License information](#) ►

### **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.



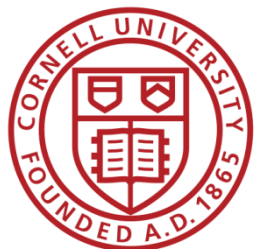
# Risk



**404.** That's an error.

The requested URL /a\_cool\_website was not found on this server. That's all we know.





# 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

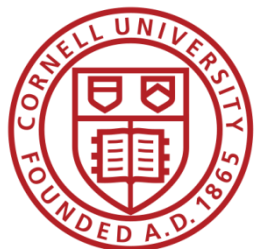
Press for the NBER; 2009. pp. 149–230.

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.




# Data not attached to article

- J of Econometrics Data Policy at the time could not accommodate 50MB file
  - Data was not attached to paper.
- Today's J of Econometrics policy suggests using third-party repositories
  - We will get to that later



# 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

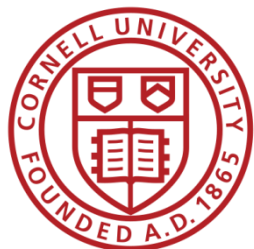
**Replication data for: National estimates of gross employment and job flows from the Q Indicators with demographic and industry detail**

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 occupation and sex), economic industry (NAICS industry groups), and detailed geography (county, Census of Economic Geography, 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

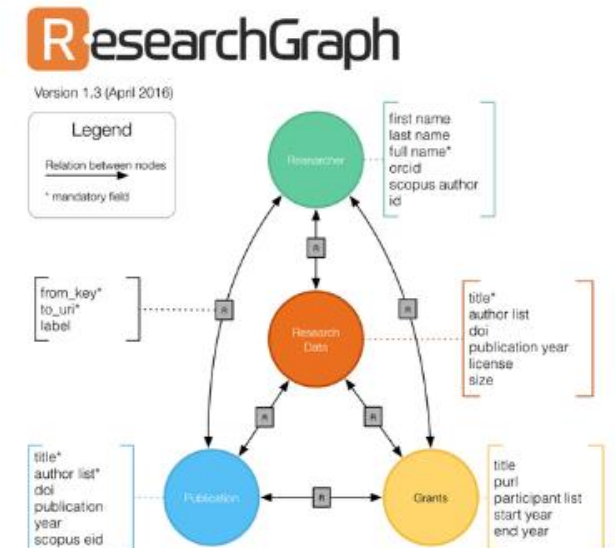
Labor Dynamics Institute (Cornell University) (LDI) <http://www2.vrdc.cornell.edu/news/data/qwi-national-data/>



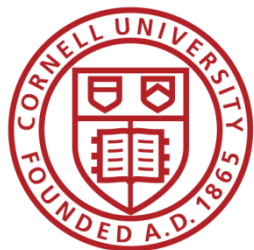


# But journal and data infrastructure are incomplete

- While Dataverse allows to manually link back...
- ... the article itself (journal website) reveals **none** of that
- True for most journals, and most data archives
  - ICPSR (manual linking to articles)
  - RePEc (no linkage possible)
- Infrastructure starting to emerge
  - If article cites data (DOI!)
  - If archive and/or journal leverages infrastructure







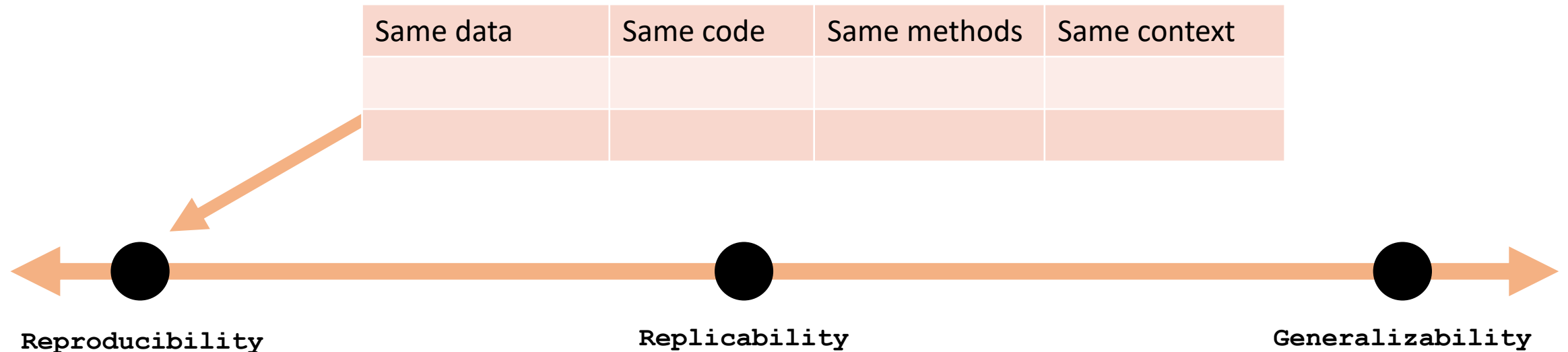
Still true today...

# Issues

Not enough  
articles are  
reproducible



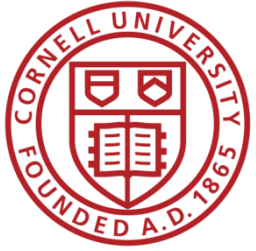
# Replication continuum



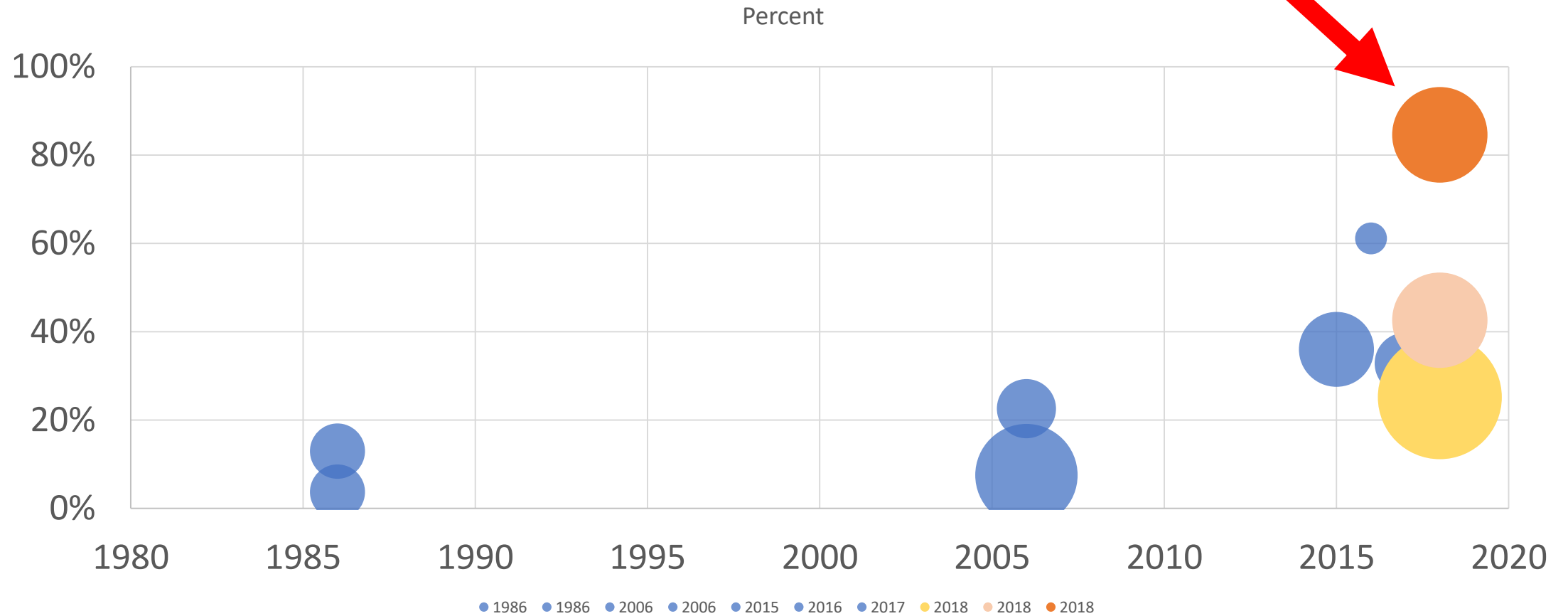
- Narrow Replication (Pesaran 2003)
- Pure Replication (Hamermesh 2007)
- Verification (Clemens 2015)

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

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



# 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.



# 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!



# Issues



Not enough  
data is  
“accessible”



# Current Data Availability Policies are Broken

- If the Data is  
**not open-access,**  
  
**no systematic information is  
collected**  
(“exemption”)

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”

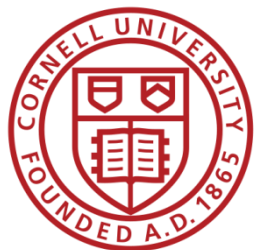


# Current Data Availability Policies are Broken

- If the Data is  
**open-access,**

**no systematic information is  
collected**

(not cited, lack of information)



What to do about it?

# Evolving Journal and Data Infrastructure





# Why do journals like “supplemental ZIP files” and affiliated repositories?

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



# What are the characteristics of prominent data archives?

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



# Evolving Journal and Data Infrastructure

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



# 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!





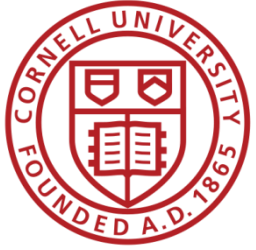
# Evolving Journal and Data Infrastructure

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

- Dataverse
- Figshare
- openICPSR
- Zenodo

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

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



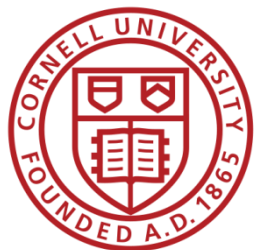
# Evolving Journal and Data Infrastructure

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



## 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)*



# Evolving Journal and Data Infrastructure

Treat all archives  
symmetrically!





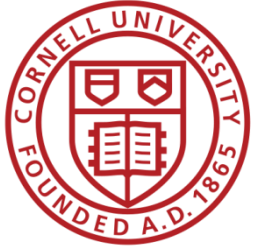
# Verifying Data and Code Deposits

- 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 restricted-access 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?**



# Evolving Journal and Data Infrastructure

**So: Describe them!**

(cite them!)



# Data (and Code) Availability Statements

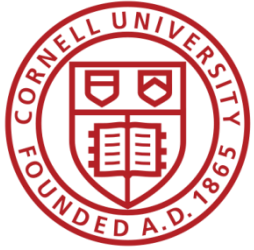
- A statement about where data supporting the results reported in a published article can be found
  - including unique identifiers linking to publicly archived datasets analyzed or generated during the study.
- DASs can increase transparency by providing a reason why data cannot be made (immediately)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



# “Unscripted Segway”

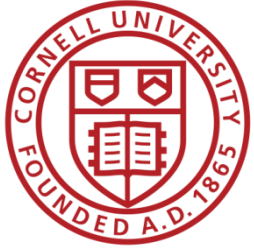
- Does that mean I cannot use data from Firm ABC?
- Does that mean I have to give my data away?

# Current efforts at the AEA



# Current efforts at the AEA

- Provide more transparency
  - To assist replication efforts
  - By better linking to paper-related resources
    - Public-use data
    - Restricted-access data
    - Code
    - Pre-Registration when available



# Current efforts at the AEA

- Pre-emptively improve code archives
  - By conducting reproducibility checks when we can
  - By working with groups that conduct reproducibility checks when we cannot





# Current efforts at the AEA

- Better archives
  - Greater transparency of the code and data archives
  - Better provenance tracking
    - Leave code where it is when appropriate
    - Leave data where it is almost always
    - Display that information



# 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” (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 theoretical, empirical, or experimental work must **provide** a **README** and data appendices. **We will assess, by reviewing** these materials, **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 are **clearly and precisely** documented and made available to the researcher for purposes of replication.
- Authors of accepted papers that contain empirical work, simulations, or experimental work must **provide, prior to acceptance, evidence that** 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.

**We will assess early,  
but more flexible.**



# 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 theoretical work, simulations, or experimental work must **provide, prior to acceptance, evidence that** 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.

“Sufficient” = “it actually works”.



# 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 academic work, simulations, or experimental work must provide **Better repository** **acceptance**, evidence that the data, programs, and computations **sufficient to permit replication**. These should be **available on the AEA Data and Code Repository**. 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 empirical work, simulations, or experiments must provide **... or other repository** **acceptance**, evidence that the data, programs, and other details of the computations **sufficient to permit replication**. These should be **available on the AEA Data and Code Repository or another repository**. 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.





# Action: 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)*

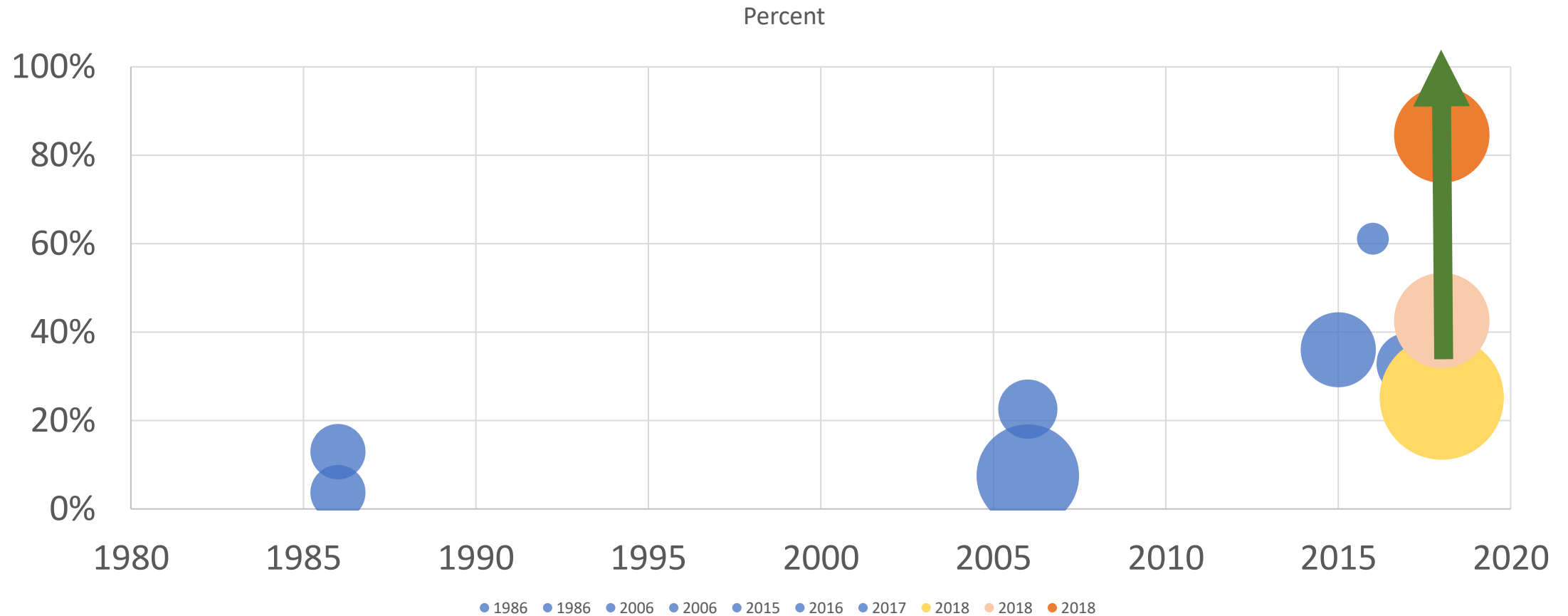


# Action: Pre-Publication Verification

- Cornell Replication Lab has been verifying published articles
- Now switching to manuscripts in the submission workflow
- For now pilot
  - *Authors have submitted prior to announcement of new data policy*



# Goal: Improve reproducibility





# Action: Verifying Data and Code Deposits

- Check README
  - Legible? Intelligible? Complete?
- Check Code
  - Where is Table 1? Figure 1? Could this work?
- Check Access Rights
  - Can the author provides us with data?
  - Does the data access as described work?

Challenges?



# 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

Impossible?



# Lots of good examples

- Open source software has practices that ensure reproducibility, but also describe it
- Many papers do an admirable job, and teach the replicator how to proceed
- Tools:
  - Make files and modern replacement
  - Docker
  - Rmd files
  - Maybe Jupyter notebooks (they have some issues)



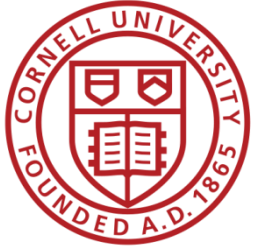


# Lots of good examples (restricted-access data)

- IAB FDZ enforces reproducibility through its access procedures
  - So does the CDER/Statistics Canada
- Some European agencies have excellent data documentation
  - So does (sometimes) Statistics Canada
- Access procedures are often quite formal but impartial
  - US, Canada, France, Germany, etc.



Lots of bad examples too....



# Future efforts

AEA, Social Sciences, elsewhere



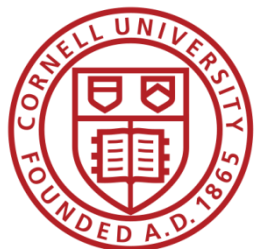
# Better support for researchers

- 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




# Moving away from “supplemental data”

- Data as a primary object
  - Title
  - Permanent location
  - Citable!
- Better data repositories
  - Move away from ZIP files attached to web pages
- Greater clarity about locations



# Full-featured repository

**OPEN ICPSR** [Find Data](#) [Share Data](#) [openICPSR Repositories ▾](#)  [GO](#) [Sign Up](#) [Sign In](#)

 **AMERICAN  
ECONOMIC  
ASSOCIATION** [AEA Deposit Instructions](#) [Browse AEA Deposits](#) [Contact](#)

---

## Depositing Data in the AEA Data and Code Repository

The *American Economic Association journals* require authors to deposit data and materials with a community-recognized or general repositories. The *AEA Data and Code Repository at ICPSR* serves that purpose. Please see the AEA's [Data and Code Availability Policy](#) and data citation guidance at the [Sample References](#) page for more details. **Authors are required to include a citation pointing to the deposit in the reference section of the final version of the article sent to the AEA.** The *openICPSR* repository automatically generates a citation when the data are "published."

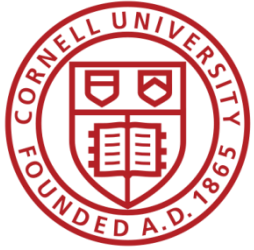
Deposits should include all data, annotated program code, command files, and documentation that is needed to replicate the findings from the authors' submitted article.

- **Data** should be comprehensively documented (see ICPSR's [Guide to Social Science Data Preparation and Archiving, 5th Edition](#) for guidance). The **author** is responsible for removing identifying information from the data to protect [confidentiality](#). Neither the AEA nor ICPSR review submissions for disclosure risk.
- **Program** code and command files should be annotated to facilitate replication and ensure clear correspondence between code and figures, tables, and analyses in the published article.
- Authors retain ownership and copyright to the data and code. Authors are required to affirm that they have the right to publish and redistribute the material. However,
  - ICPSR requires a license for distribution of data.
  - An **open license** is required by the AEA, in order to allow others to re-use the data and code, in particular for replication. Authors can select from several license options, including CC-BY 4.0 for data and Modified BSD for software and code. If an author would like to use multiple licenses or create a customized license, she should select the "Other" license option and upload a LICENSE file alongside the data and documentation.

By depositing in the AEA Data and Code Repository, the depositors allow the AEA staff to add keywords and other metadata which are important for proper indexing in linking. Any other changes are subject to the license chosen for the materials.

[View more extensive \(unofficial\) guidance.](#)

Start Your Deposit



# Full-featured repository

OPEN ICPSR

[Find Data](#)

[Share Data](#)

[openICPSR Repositories](#) ▾



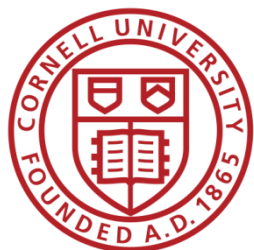
AMERICAN  
ECONOMIC  
ASSOCIATION

## Search Results

[search tips](#) ▾

**No studies were found that match your criteria.**

- Try broadening your search and removing filters.
- Check your spelling.
- Still no luck? Email [help@icpsr.umich.edu](mailto:help@icpsr.umich.edu) or [suggest a data collection](#) that we should acquire.



# Predation, Protection, and Productivity: A Firm-Level Perspective.



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 Association* [publisher] DOI: [10.1257/mac.20160120.data](https://doi.org/10.1257/mac.20160120.data)

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

- 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)

- ☐ Data restricted-access, under Norwegian Data Access license (has residency requirement, has citizenship requirement), accessible at [Norwegian Data Archive in Oslo, Norway](#)



# 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

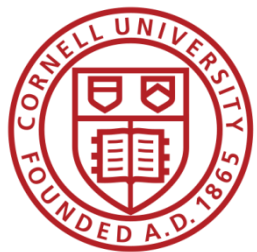
See [Verification guidance](#)

### FAQ

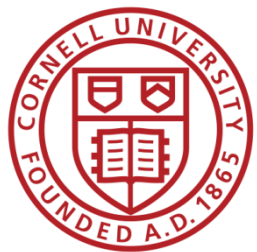
See our growing [FAQ](#). If you have questions or answers to add, please notify us by creating a [new issue](#).

[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...



FileEditViewRepositoryBranchHelp

Current repository  
replicability-presentation2019

Current branch  
master

Pull origin

Last fetched just now

1

ChangesHistory

0 changed files

No local changes

You have no uncommitted changes in your repository! Here are some friendly suggestions for what to do next.

Pull 1 commit from the origin remote

The current branch (master) has a commit on GitHub that does not exist on your machine.

Always available in the toolbar when there are remote changes or **Ctrl Shift P**

Pull origin

Open the repository in your external editor

Configure which editor you wish to use in [options](#)

Repository menu or **Ctrl Shift A**

Open in Atom

View the files in your repository in Explorer

Repository menu or **Ctrl Shift F**

Show in Explorer

Open the repository page on GitHub in your browser

Repository menu or **Ctrl Shift G**

View on GitHub

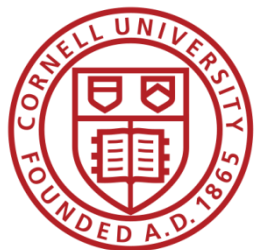
Summary (required)

Description

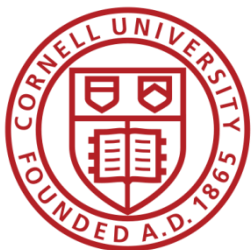
1+

Commit to master





New skills to learn...

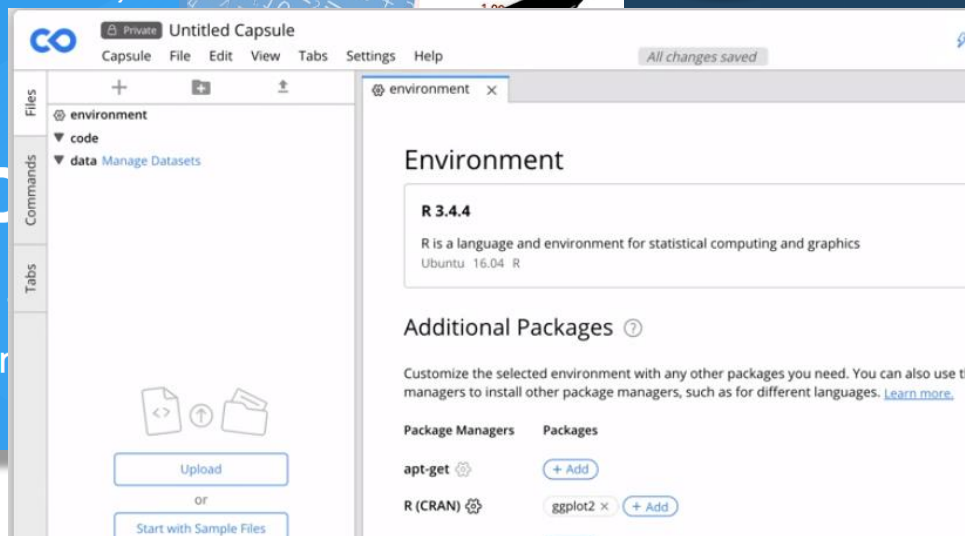


## Shape Your D

Join us at DockerCon 2019,  
things Kubernetes, microser

## R Markdown

from R Studio



About Downloads Documentation

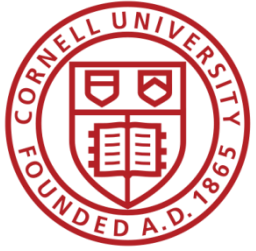
```
n 3: Fibonacci series up to n
fib(n):
a, b = 0, 1
while a < n:
    print(a, end=' ')
    a, b = b, a+b
```

(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
```

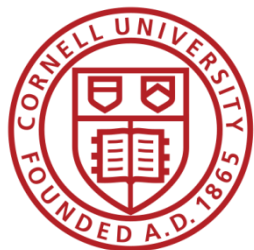
```
>>> r = requests.get("https://zenodo.org/api/deposit/depositions")
>>> r.status_code
401
>>> r.json()
```

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



We!

- Ingrained habits
- New skills to learn
- New methods to use



Push for better support...

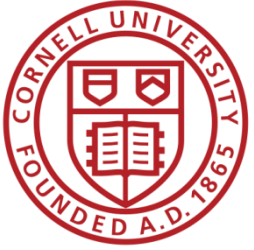


# Researchers: New skills to learn/teach

- 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!

# Glimpses





# Some random notes

- Analogy between **grant** or **RDC proposal** and **pre-registration**
- Incentives of stats agencies: **transparency** = **credibility**
- Challenges with **ad-hoc access** (individuals accessing ministry data, CD in the back pocket/file drawer, unnamable private company)
- From **pre-acceptance verification** to **pre-submission verification** (university or institute services) and the role of **contract programming**

# Summary



# Goals

- **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



# Challenges for Restricted-Access Data

- **Verifiability**

- How can others obtain access?

- **Documentation**

- How can others learn about the data?

- **Persistence**

- How are data and programs preserved?

**Thank you!**

**DOI: 10.5281/zenodo.2573123**