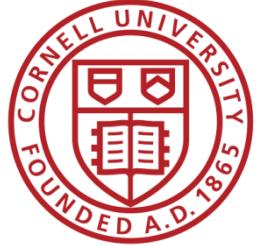


Replication and Reproducibility in Social Sciences and Statistics: Context, Concerns, and Concrete Measures when Data are Confidential

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.



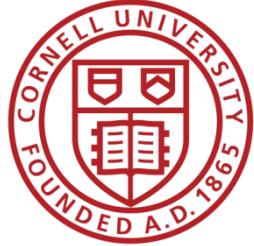
Disclaimer

My opinions

Not Census Bureau

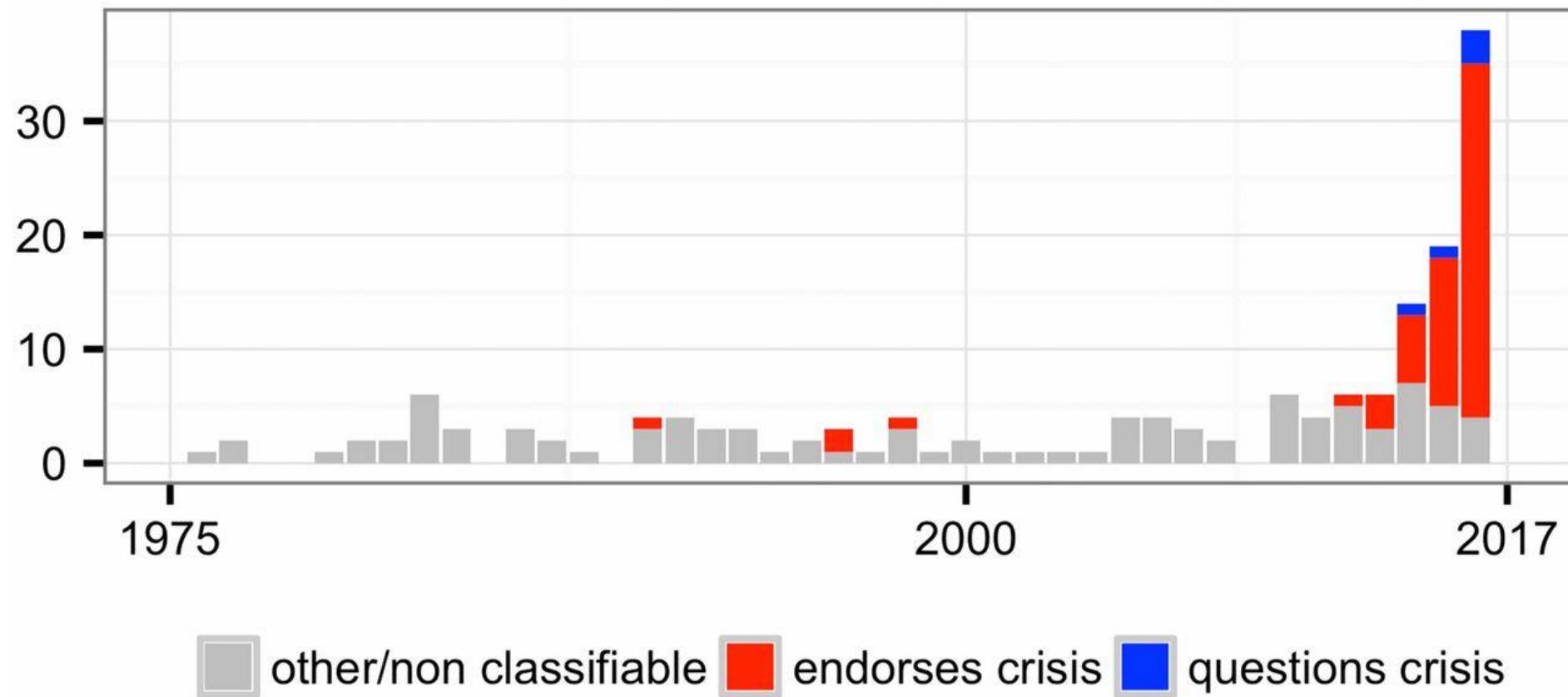
Not American Economic Association

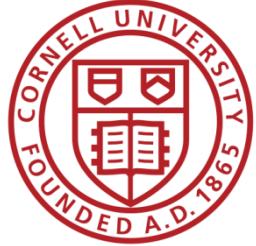
No confidential data was abused in this talk



This reproducibility crisis thing....

Frequency of Crisis Narrative in Web of Science Records





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

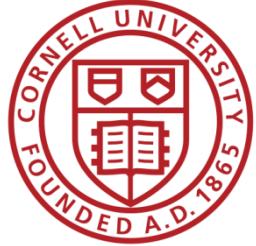
Selective reporting

Ignoring nulls

Lack of replication

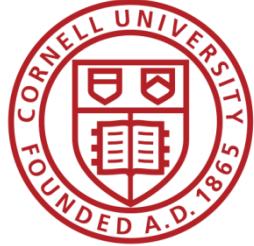
Misuse of statistics

Source: Nosek
Sackler talk 2017



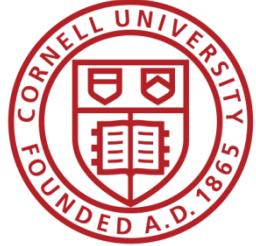
Efficiency of scholarly discourse?

- 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**



SEASONAL VARIATIONS IN THE NEW YORK MONEY MARKET, 1890-1908																	
CALL INTEREST RATES ON STOCK EXCHANGE ^b				INTEREST RATES ON 60-90 DAY, 2 NAME COMMERCIAL PAPER ^b		PERCENTAGE OF RESERVES TO DEPOSITS, N. Y. ASSOCIATED BANKS ^b		CIRCULATION OF DEPOSIT CURRENCY ^b		EXCHANGE RATES IN CHICAGO ON NEW YORK, ^c 1899-1908		NET INTERIOR MOVEMENT OF CASHS OUT OF AND INTO N. Y. CITY BANKS, ^c 1899-1908		STERLING EXCHANGE, DEMAND DRAFFTS ^f		EXPORTATION AND IMPORTATION OF GOLD, U. S., 1890-1908 (IN FIGURES) ^e	
AVERAGE RATE	SEASONAL INDEX NUMBER	AVERAGE RATE	SEASONAL INDEX NUMBER	AVERAGE PERCENTAGE	SEASONAL INDEX NUMBER	AVERAGE CLEARINGS (\$000,000)	SEASONAL INDEX NUMBER	AVERAGE RATE (PREMIUM OR DISCOUNT)	SEASONAL INDEX NUMBER	AVERAGE AMOUNT OUT OF 000	INTO 000	SEASONAL INDEX NUMBER	AVERAGE RATE	SEASONAL INDEX NUMBER	TOTAL EXPORTS 000	TOTAL IMPORTS 000	
6.4	43.4	5.0	53.1	28.6	44.3	\$1,237.5	60.8	2.5 P	64.7	86,084	87.2	48.606	49.7	Jan.	832,747		
3.6	23.8	4.7	41.5	29.1	78.8	*1,233.6	*59.6	5 P	67.4	6,621	84.9	48.657	54.7				
2.8	14.9	4.5	31.2	29.9	96.9	*1,234.7	*54.4	5 P	67.7	7,773	90.7	48.679	59.4				
2.5	11.9	4.3	22.7	30.3	77.8	*1,140.0	*44.0	10 P	72.1	6,895	87.6	48.697	64.1				
2.5	11.1	4.3	22.9	29.9	65.4	*1,190.5	*52.5	2 P	63.0	4,749	64.1	48.695	64.1	Feb.	13,408		
2.4	10.1	4.3	22.1	29.2	58.1	*1,084.1	*38.4	6 D	54.8	2,376	77.0	48.696	64.8				
2.5	9.8	4.3	22.2	28.8	53.6	*1,004.8	*32.1	9 D	50.7	1,436	63.7	48.708	66.9				
2.7	13.4	4.4	26.5	28.5	53.6	*944.0	*22.6	20 D	38.8	1,157	52.5	48.697	65.4				
3.0	15.1	4.6	32.6	28.1	45.5	*1,165.7	*51.5	29.5 D	28.1	1,679	58.5	48.692	65.7	March	\$ 43,233		
3.6	19.7	*4.6	*34.3	27.9	43.1	*1,067.9	*38.2	23 D	35.0	604	30.5	48.676	62.0				
3.9	22.4	4.8	40.0	27.7	37.0	*1,119.7	*42.7	13 D	45.9	716	49.8	48.665	59.1				
3.2	19.2	4.8	39.6	27.9	39.9	1,042.3	33.1	14.5 D	43.5	1,533	54.4	48.681	61.6				
3.6	22.0	4.8	38.1	28.0	40.5	1,051.4	35.5	5 D	53.9	999	53.5	48.704	65.9	April	25,888		
4.0	23.8	4.7	36.7	27.8	35.7	1,135.4	48.0	14 D	44.5	868	53.9	48.711	67.4				
3.8	23.1	4.6	33.4	27.9	39.9	1,119.0	42.9	7.5 D	52.9	1,903	59.0	48.714	68.2				
3.0	17.5	4.5	31.9	28.4	50.9	1,123.5	46.7	4 P	66.3	2,085	62.1	48.734	73.6				
2.9	15.4	4.4	27.5	28.6	54.4	1,107.6	43.3	9 D	48.4	1,379	61.6	48.743	78.1	May			
3.4	19.3	4.4	26.9	28.3	48.3	1,283.3	67.3	3.5 D	55.9	594	56.5	48.739	76.3				
3.5	19.5	4.4	24.5	28.4	48.0	1,175.4	52.7	2.5 P	62.0	9,952	63.0	48.734	74.2				
2.6	13.9	4.3	22.7	28.6	51.6	1,123.4	48.0	16 P	76.7	4,306	74.5	48.739	75.5				
2.4	11.2	4.2	19.9	29.0	60.3	1,011.8	34.1	16 P	77.3	4,329	74.7	48.752	79.1	June			
2.3	9.6	4.1	17.1	28.8	57.2	908.1	21.4	10 P	71.1	3,862	60.9	48.760	80.9				
2.3	8.0	4.1	15.8	28.7	56.1	1,039.4	37.9	5 P	64.6	3,229	68.6	48.757	81.1				
2.4	7.7	4.1	15.3	28.7	56.7	967.8	31.1	4 P	63.6	3,354	66.7	48.756	81.0				
2.5	8.0	4.3	18.4	28.7	57.5	938.7	25.8	10.5 P	72.8	3,897	68.5	48.742	79.0	July			
3.6	16.4	4.5	22.0	28.4	53.5	1,013.9	35.4	11.5 P	73.6	2,158	58.3	48.721	74.6				
3.4	13.6	4.5	25.0	27.9	45.0	991.5	33.1	16.5 D	40.3	1,441	53.1	48.715	72.9				
2.9	9.6	4.6	26.9	28.4	56.3	1,034.6	35.6	7.5 D	50.6	3,456	68.0	48.717	72.6				
2.3	5.3	4.6	31.1	28.7	63.3	970.2	26.6	8 D	52.6	3,692	69.3	48.717	72.6				
2.4	5.6	4.6	33.5	28.7	65.4	924.6	21.1	10.5 D	50.0	4,735	73.1	48.720	73.2				
2.5	6.0	4.6	35.2	28.3	60.8	969.7	27.9	11 D	48.7	2,955	63.4	48.702	69.6	August	44,300		
2.5	6.3	4.8	40.5	28.0	54.3	910.6	20.8	17.5 D	41.8	1,395	57.3	48.693	68.0				
2.6	7.4	4.9	43.7	27.8	49.3	948.0	25.9	19 D	40.1	9,517	49.4	48.669	61.3				
3.7	13.6	5.3	49.5	27.7	47.7	931.1	23.9	34.5 D	92.7	8249	45.5	48.651	56.9				
3.0	12.3	5.3	51.8	27.6	42.6	956.8	29.0	37.5 D	18.8	1,477	33.7	48.626	50.4				
4.1	20.7	5.3	55.4	27.2	32.8	880.7	19.2	36.5 D	19.1	2,690	29.9	48.601	43.7				
4.2	23.4	5.1	57.5	27.0	29.8	1,033.6	38.6	25 D	34.7	2,589	30.3	48.584	35.2				
4.3	30.6	5.3	64.7	27.1	31.9	1,058.7	44.3	26 D	33.5	3,434	34.8	48.552	32.0				
4.2	29.6	5.3	63.2	27.5	37.4	1,066.1	36.9	33 D	26.1	3,489	37.0	48.557	31.9	Oct.			
4.5	27.9	*6.2	*61.7	27.3	33.0	1,135.2	59.0	32 D	27.2	3,883	39.0	48.538	27.3				
4.0	24.4	*5.1	*61.5	27.3	33.0	1,094.1	46.4	29.5 D	29.0	32.5	30.3	48.540	29.7				
3.6	19.4	*4.9	*53.2	27.5	34.1	1,139.3	49.6	27.5 D	30.8	3,014	34.7	48.549	33.9				
6.5	29.3	*4.9	*51.4	27.6	36.4	1,144.0	50.1	31 D	24.2	3,685	34.7	48.576	41.5				
7.1	32.9	*4.9	*48.9	27.2	27.5	1,140.7	54.3	29 D	27.5	2,700	37.1	48.567	39.7	Nov.	96,743		
5.4	30.3	*4.9	*51.3	27.1	29.7	1,077.6	45.3	20 D	36.9	2,666	43.6	48.554	38.8				
4.8	26.1	*5.0	*53.5	27.4	29.4	1,983.9	65.7	4.5 D	33.4	1,530	48.6	48.594	44.1				
4.2	26.1	*4.7	*46.0	27.8	36.1	1,107.7	48.1	2.5 D	56.3	49.0	48.623	49.5					
4.0	26.8	4.8	48.6	27.6	32.3	1,191.3	65.2	11.5 D	47.3	836	44.3	48.615	49.3				
4.9	30.3	*4.7	*47.8	27.2	24.9	1,202.4	63.5	5 P	64.7	615	52.4	48.596	45.6				
5.5	39.2	*4.8	*51.6	27.4	29.4	1,202.1	60.8	3.5 P	65.1	60	47.8	48.604	47.0				
6.6	46.1	*4.8	*49.3	27.5	32.8	1,015.3	35.9	3.5 P	65.1	9,188	61.7	48.611	49.0				
7.4	49.3	*4.9	*52.9	27.7	35.3							48.592	45.0				

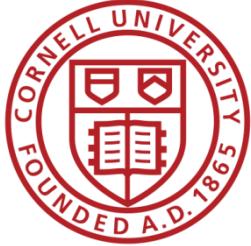
From @sdellavi
AER 1911



Efficiency of scholarly discourse!

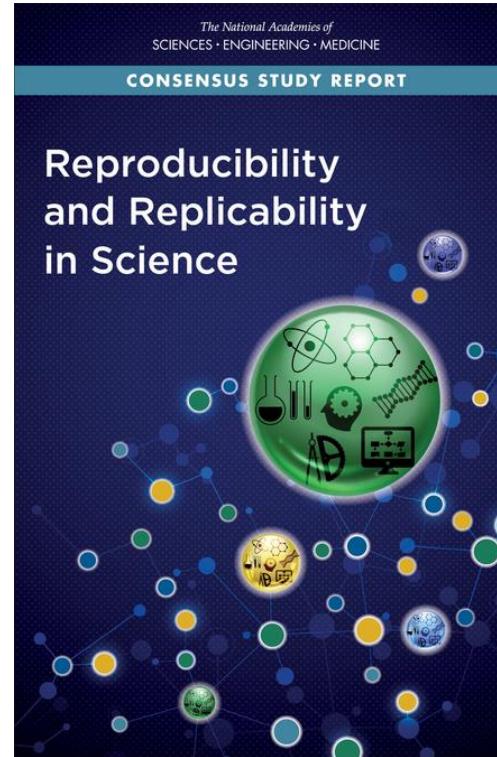
**Modern publications thus need
the same transparency and completeness
as in the old days
to facilitate replicability**

Replication?



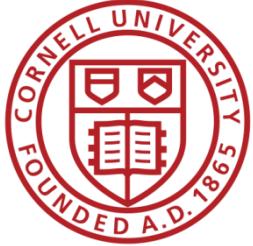
Replication continuum

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

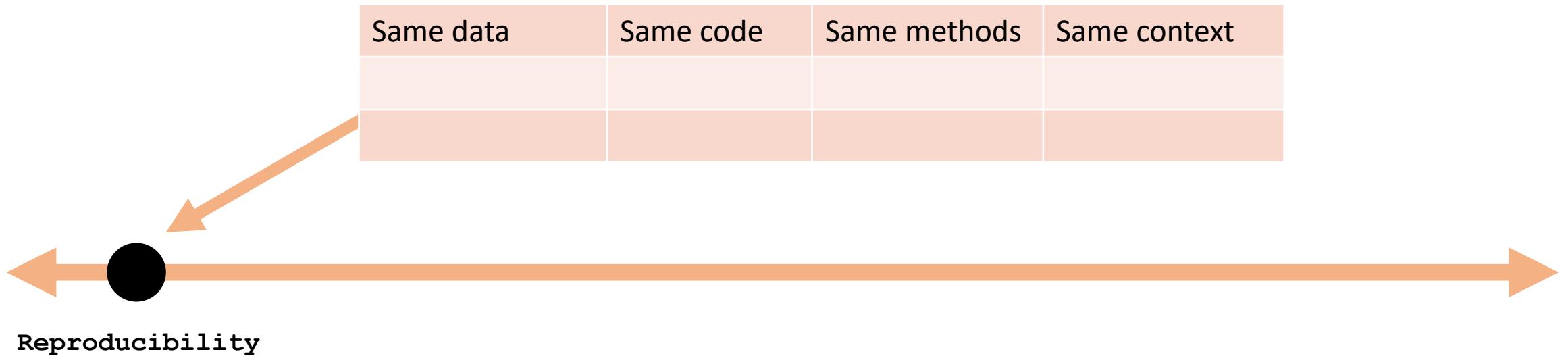


Reproducibility

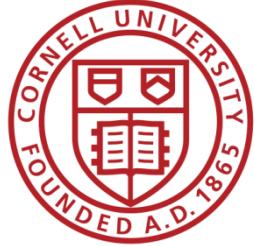
- Narrow Replication (Pesaran 2003)
- Pure Replication (Hamermesh 2007)
- Verification (Clemens 2015)



Replication continuum



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



Replication continuum

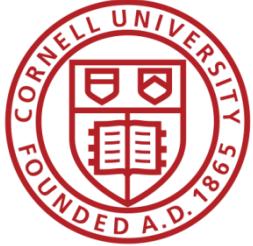


Reproducibility

- 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

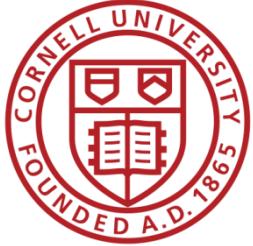


Reproducibility

- 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

New data collection	Same code	Same methods	Same context

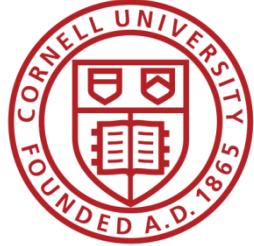


Reproducibility

- 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



Reproducibility

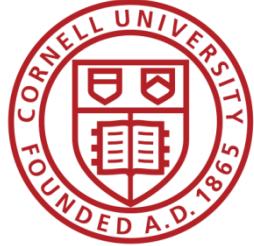
- Narrow Replication (Pesaran 2003)
- Pure Replication (Hamermesh 2007)
- Verification (Clemens 2015)

Replicability

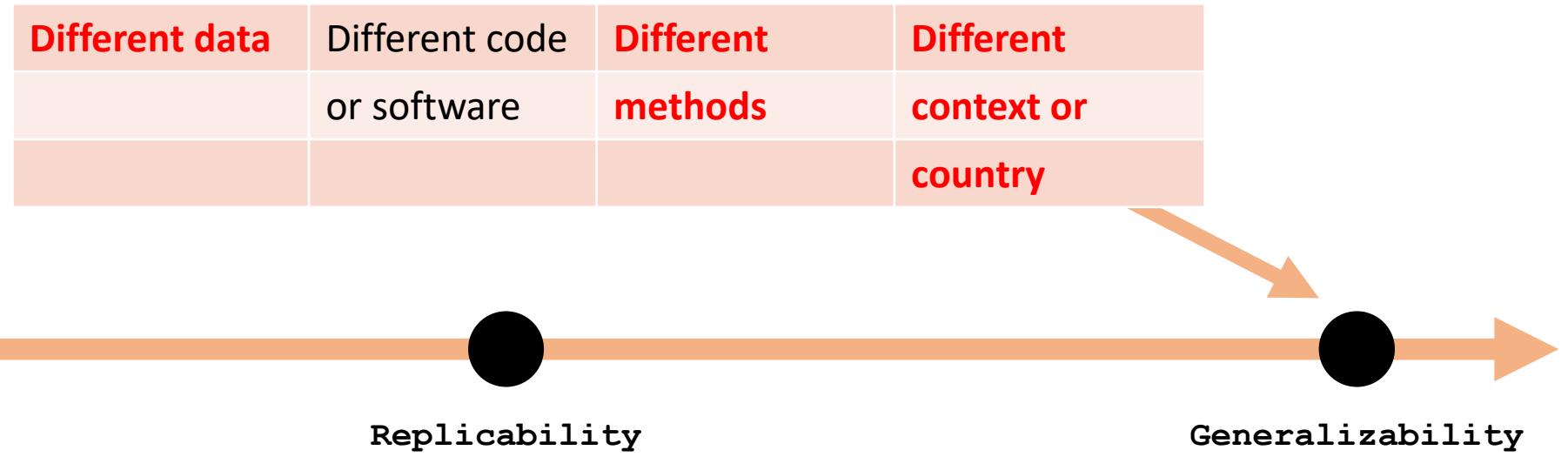
- Wide Replication (Pesaran 2003)
- Statistical Replication (Hamermesh 2007)
- Reproduction/Reanalysis (Clemens 2015)

Generalizability

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



Replication continuum

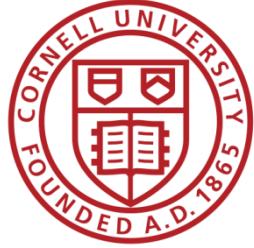


- Narrow Replication (Pesaran 2003)
- Pure Replication (Hamermesh 2007)
- Verification (Clemens 2015)

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

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

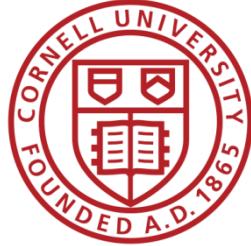
Progress



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@bc.edu)

[Access Statistics](#) for this software series.

Track citations for all items by [RSS feed](#)

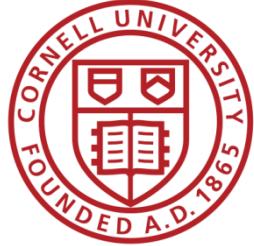
Is something missing from the series or not right? See the RePEc data [series](#).

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

Ulrich Kohler

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

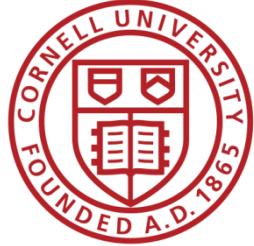
[GPROD: Stata module to extend egen for product of obs](#)
Philip Ryan



Progress

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



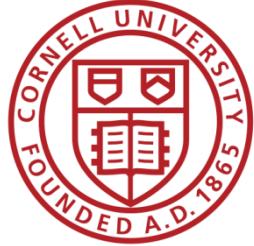


Progress

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

RePEc



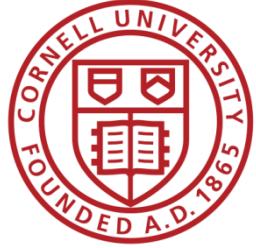


Progress

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

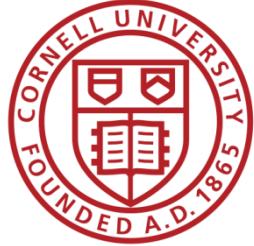


More
recently...



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)

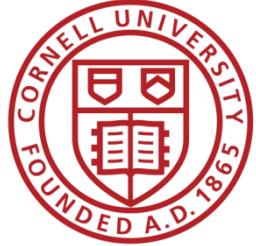


Pre-registration

- “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**.”

A screenshot of the AEA RCT Registry website. The header includes the logo for the American Economic Association and the text "AEA RCT Registry - The American Economic Association's registry for randomized controlled trials". Below the header are links for "About", "Registration Guidelines", and "FAQ". There is also a search bar and a button to "REGISTER A TRIAL". The main content area displays two registered trials:

- Tackling sexual harassment Evidence from India**: Last registered on January 26, 2019. A brief description notes that Goal 5 of the sustainable development goals adopted by the United Nations in 2015 aims to eliminate all forms of discrimination and violence against women in public and private spheres and to undertake reforms to give women equal rights to economic resources and access to ownership of property. Government of India has identified ending violence against women as a key national priority too. Brutal gangrape of a 23-year-old woman in 2012 in the capital of India led to an outcry against public apathy towards endemic sexual assault and harassment against women. A UN women's study showed that 92% of women surveyed in Delhi had suffered from either sexual, visual or verbal harassment. Pervasive sexual harassment can have debilitating impacts on psychological, economic and social lives of the...
- Malleability of Sustained Attention**: Last registered on January 25, 2019. A brief description states that the economics and education literatures traditionally view human capital as an individual's stock of knowledge and skills. In this project, we posit an additional potential component: the capacity for sustained attention. In cognitive psychology, the mind's ability to direct and sustain attention is thought to underlie all activity: cognitive processes (such as solving a math problem) as well as non-cognitive activities (such as exerting self-control) (Chun et al. 2011). In this project, we examine whether the capacity for exerting sustained attention is malleable. Using a field experiment, we introduce a novel tablet-based adaptive learning platform into low-income Indian primary schools. The platform engages students in sustained practice in either mathematics or cognitive activities ...

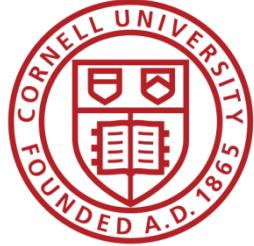


Registered Reports

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



- Close cousin: Results-blind review



Preprints in other sciences

- bioRxiv (2013)
- PsyArXiv (2016)

bioRxiv

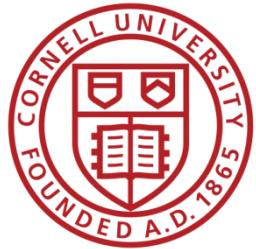
THE PREPRINT SERVER FOR BIOLOGY

Search

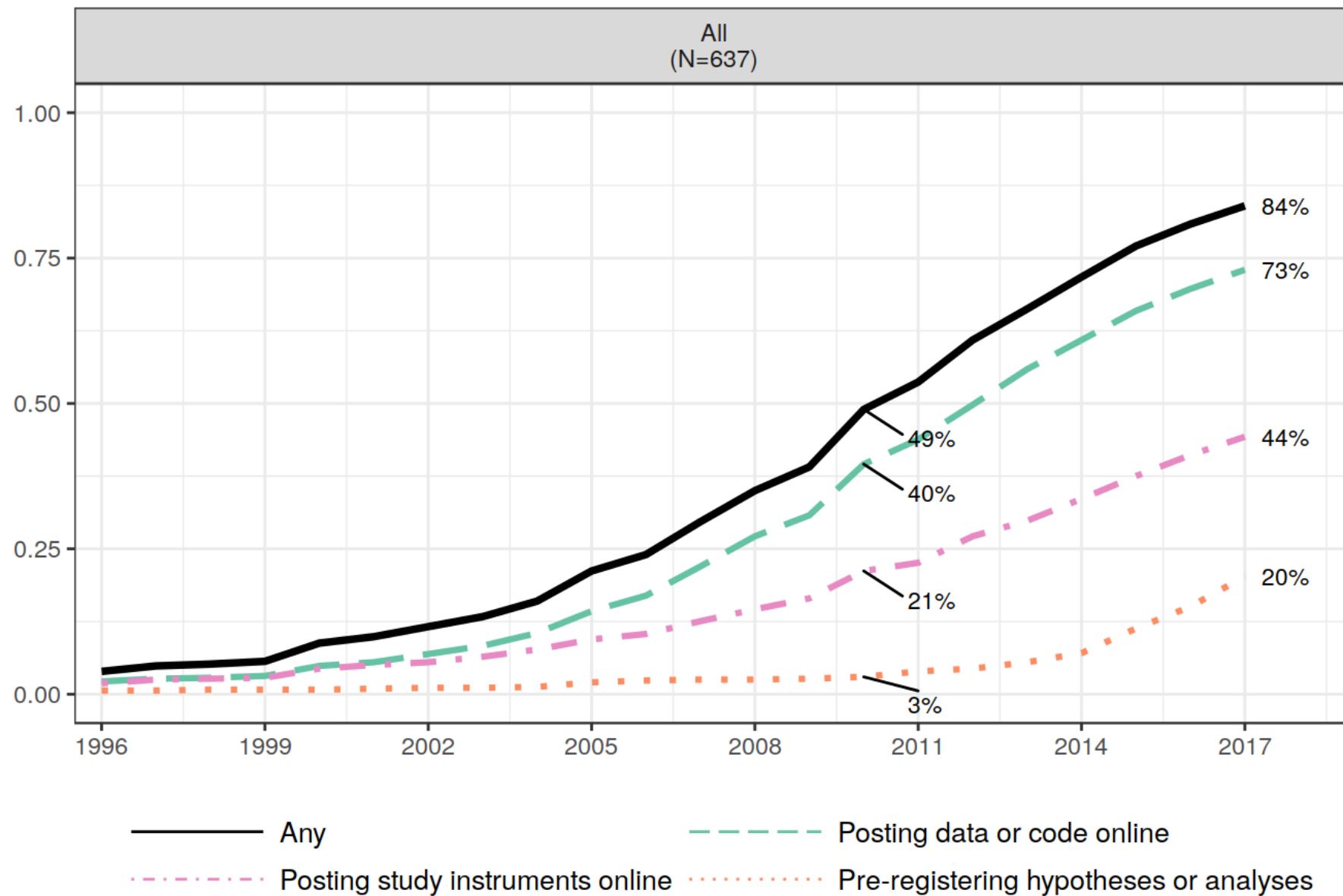


Advanced Search

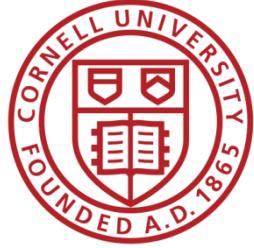




Share of Published Authors (PhD < 2010) Adopting Practice

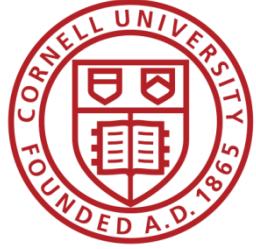


Restricted-
Access Data
Pose Problems



Challenges

- **Documentation**
 - How can others learn about the data?
- **Verifiability**
 - How can others obtain access?
- **Persistence**
 - How are data and programs preserved?

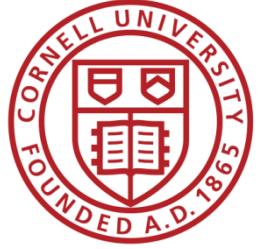


Challenges

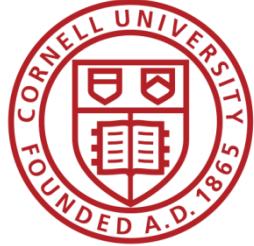
Documentation

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



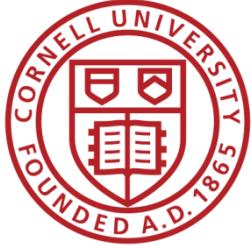
Economics makes wide use of public-use data

- **Macrodata:**

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

- **Microdata:**

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



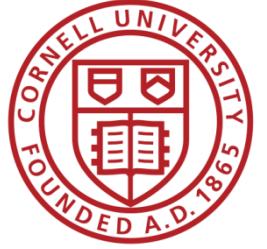
Relevant example

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

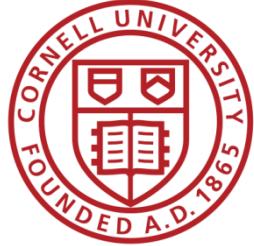
Page 5 of Online Appendix:

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.



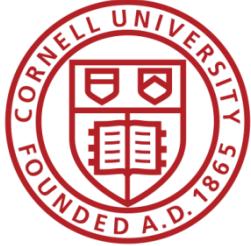
This should be easy!



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:

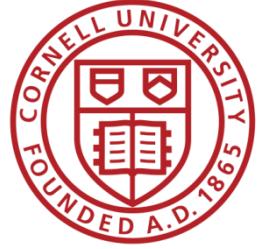
- Employment – job creation and destruction
- Job expansions and contractions
- Number of establishments
- Establishment openings and closings
- Number of startups and firm shutdowns

The BDS se

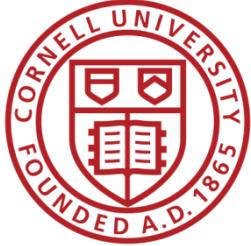
- Annu
- C
- C
- C
- Cov

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

Doesn't make it easy!



< sidenote >



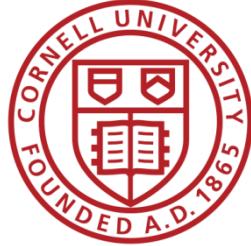
What is a data citation?

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

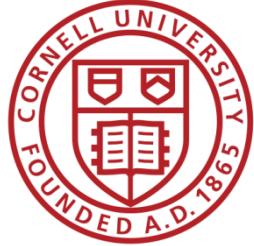


What is a data citation?

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?

- Findable,
- Accessible,
- Interoperable, and
- Re-usable

The FORCE11 logo features a blue circular icon with a white target-like pattern next to the word "FORCE11". Below it is the tagline "The Future of Research Communications and e-Scholarship". A navigation bar below the logo includes "ABOUT", "COMMUNITY", and "CODE OF CON".

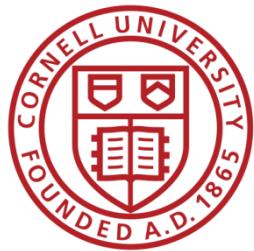
FORCE11 » Groups » The FAIR Data Principles

THE FAIR DATA PRINCIPLES

JOIN IN THE DISCUSSION - LEARN
FAIR Data Principles

Preamble

One of the grand challenges of data-intensiv



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

India [edit](#) [remove](#)

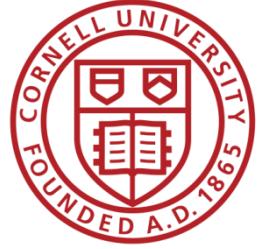
Time Period(s)

2000 – 2013 [edit](#) [remove](#)

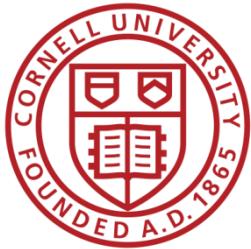
Collection Date(s)

Universe

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



</sidenote>

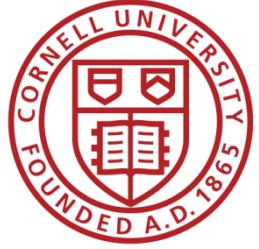


Example: BDS

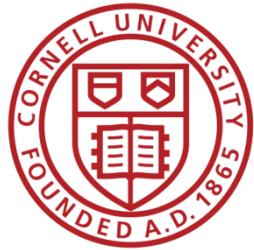
Name	Last modified	Size Description
Parent Directory		-
2015/	14-Sep-2017 09:41	-
2016/	11-Oct-2018 14:19	-
estab/	06-Sep-2016 07:50	-
firm/	06-Sep-2016 07:47	-

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)

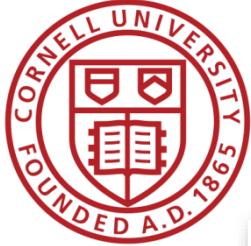
The Longitudinal Business Database (LBD) restricted-use microdata is accessible only to qualified products, the LBD provides important new insights about business formation and growth, the nature of connections to credit markets and financing to name a few. The LBD and [Business Dynamic Statistics](#) track 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 comprising

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	Jarmain, 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).

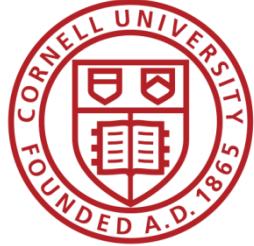


[Citation of the data in publications](#)

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:

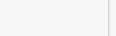
10.5164/IAB.BHP7516.de.en.v1



For example 10.1145/2783446.2783605

Select Formatting Style:

apa



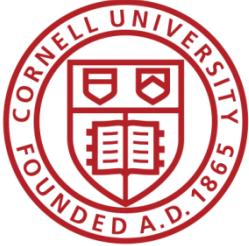
Schmucker, A., Eberle, J., Ganzer, A., Stegmaier, J., & Umkehrer, M. (2018). Establishment History Panel 1975-2016 (Version v1) [Data set]. Forschungsdatenzentrum der Bundesagentur für Arbeit (BA) im Institut für Arbeitsmarkt- und Berufsforschung (IAB). <https://doi.org/10.5164/iab.bhp7516.de.en.v1>

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 official website for the United States Census Bureau's Economic Studies (CES). The header includes the Census Bureau logo and links for TOPICS, GEOGRAPHY, LIBRARY, and DATA. Below the header, a breadcrumb trail shows the path: < Industry > Center for Economic Studies > RDC Research Opportunities > How to Apply. The main content area is titled "Economic Studies (CES)" and features a green navigation bar with links for "RDC Research Opportunities" (which is highlighted), "Research Programs", "Publications and Reports", and "FAQs". The "How to Apply" section contains instructions for researchers, a list of proposed project requirements, and a link to "Developing and Submitting a Research Proposal".

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 matter of the RDCs. All researcher access to restricted-use data occurs at secure [Federal Statistical Research Data Centers](#).

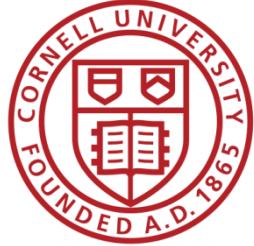
Proposed projects must

- 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/rdcr esearch/howtoapply.html>

A screenshot of the United States Census Bureau website. The header features the Census Bureau logo and navigation links for TOPICS, GEOGRAPHY, LIBRARY, and DATA. Below the header, a breadcrumb trail shows the path: < Industry > Center for Economic Studies > RDC Research Opportunities > How to Apply. The main title is "Economic Studies (CES)". A green navigation bar below the title includes links for "RDC Research Opportunities", "Research Programs", "Publications and Reports", and "FAQs". The main content area is titled "How to Apply" and contains text about proposal requirements and a list of bullet points. Another section titled "Developing and Submitting a Research Proposal" provides instructions for preparing a proposal, including a link to "Research Proposal Guidelines".

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 matter. All researcher access to restricted-use data occurs at secure [Federal Statistical Research Data Centers](#).

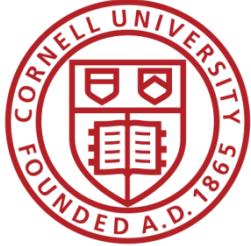
Proposed projects must

- 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)



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](#)

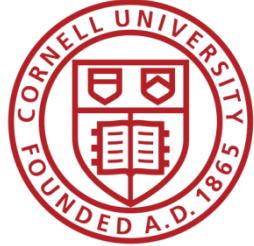
Data Access: On-site Use

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:

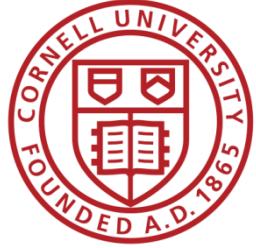
- How to obtain data access to the FDZ data via On-site Use?
- How to modify an existing FDZ agreement?
- Data use for students
- Specifics on the data access at FDZ locations in USA, Canada and UK

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 /



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

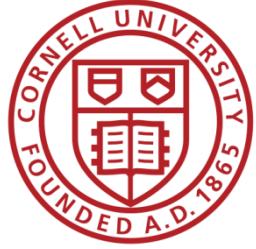
- Documentation

Verifiability

How can others obtain access?

- Persistence

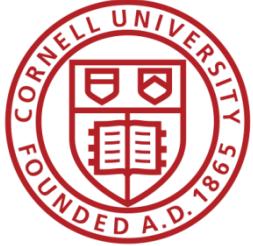
- How are data and programs preserved?



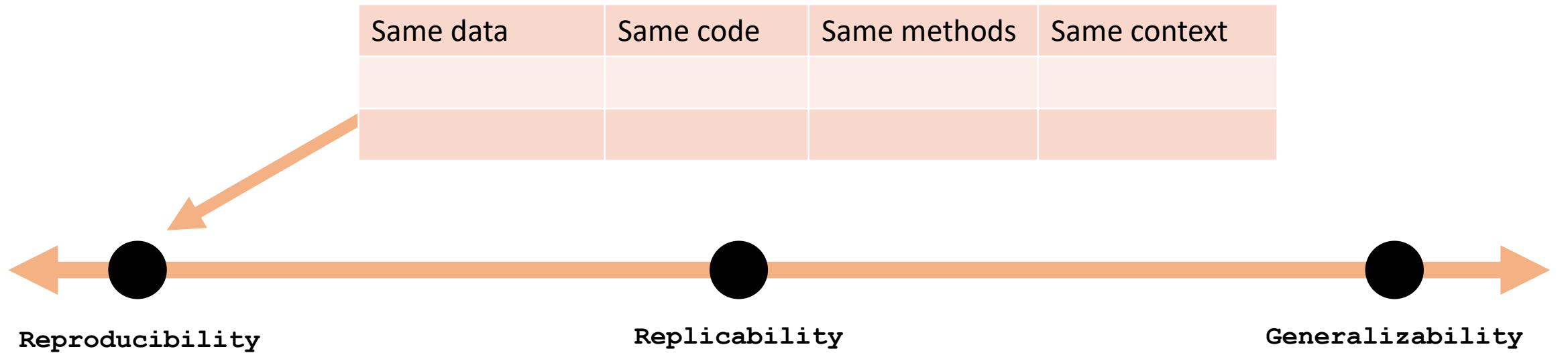
Reproducibility problems

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

Not enough
articles are
reproducible



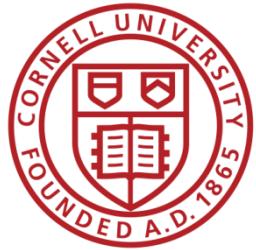
Replication continuum



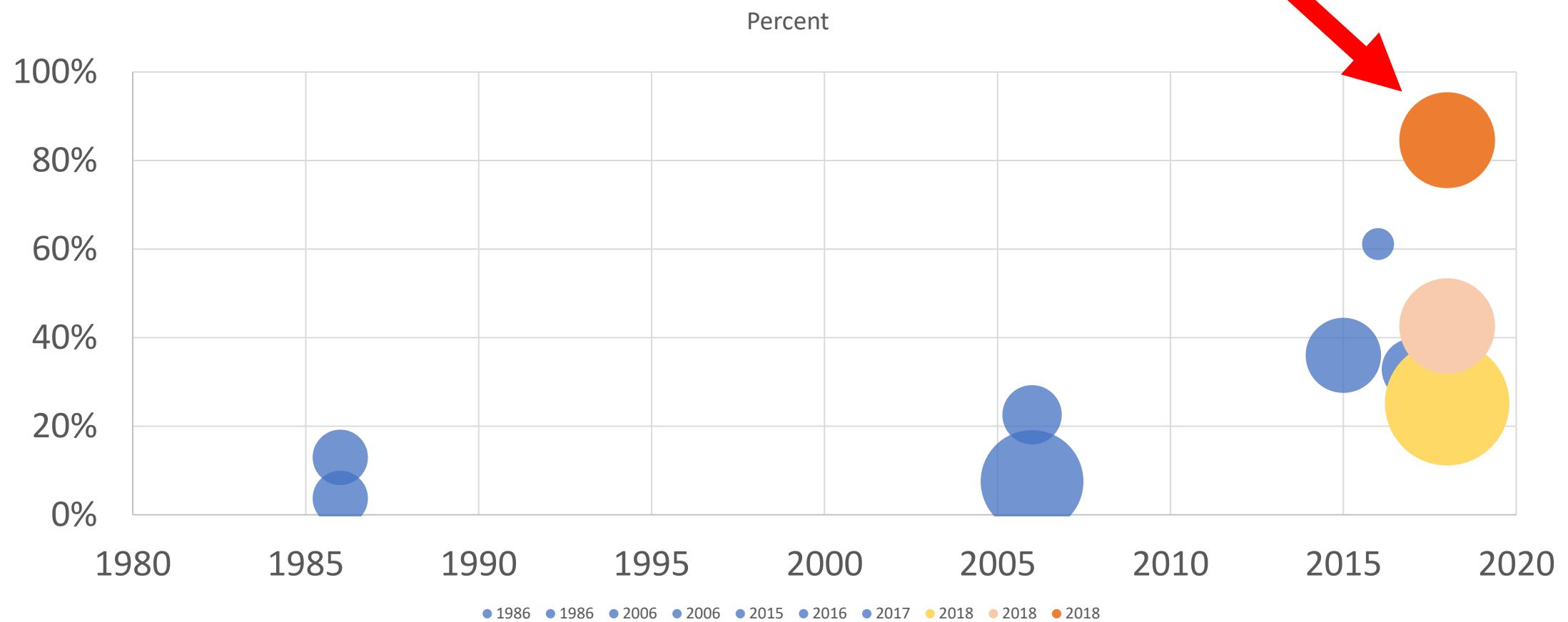
- Narrow Replication (Pesaran 2003)
- Pure Replication (Hamermesh 2007)
- Verification (Clemens 2015)

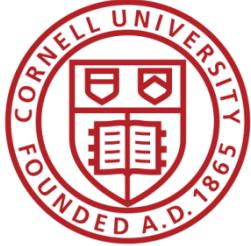
- Wide Replication (Pesaran 2003)
- Statistical Replication (Hamermesh 2007)
- Reproduction/Reanalysis (Clemens 2015)

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



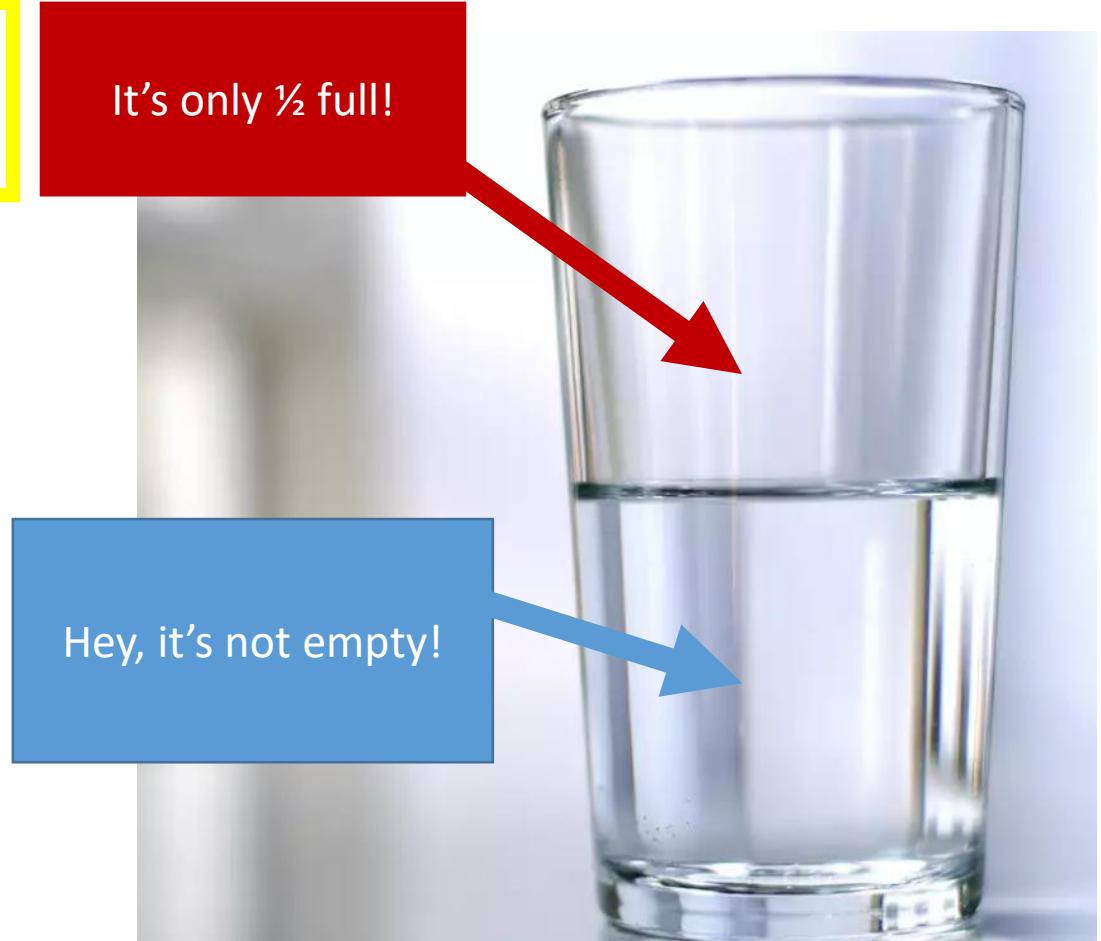
Results?

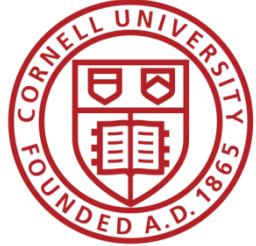




In a nutshell

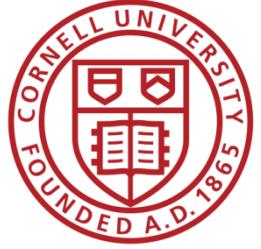
- **40%** use restricted-access data
- **25%** use public-use data and are mostly or completely reproducible
- **25%** use public-use data and are only partially reproducible
- **10%** fail to yield useful results



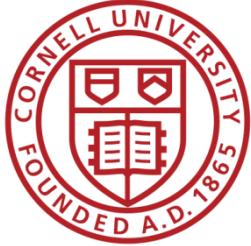


Reproducibility is harder than it should be

- 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

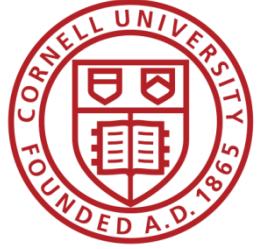


Making USEFUL archives

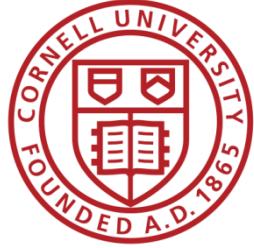
- From analysis of code from 1996 to 2003 (MMH2006):

“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.”

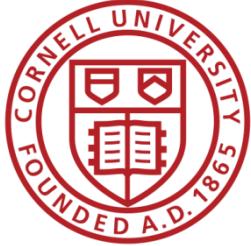


Especially true in FSRDC...



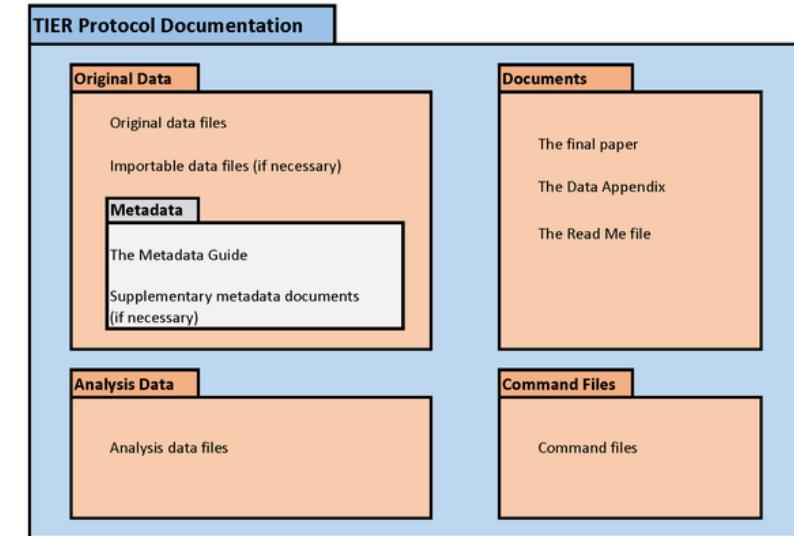
Issues in RDCs

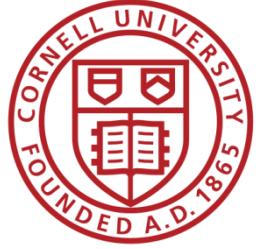
- Changes in file locations and directory
- Changes in computing architecture
- Not specific to RDCs:
 - Change in RA team
 - Data flow management
 - Broken dependencies



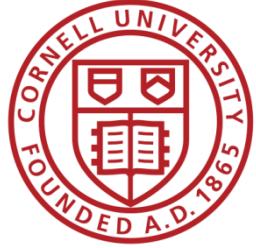
Use best practices for reproducibility

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

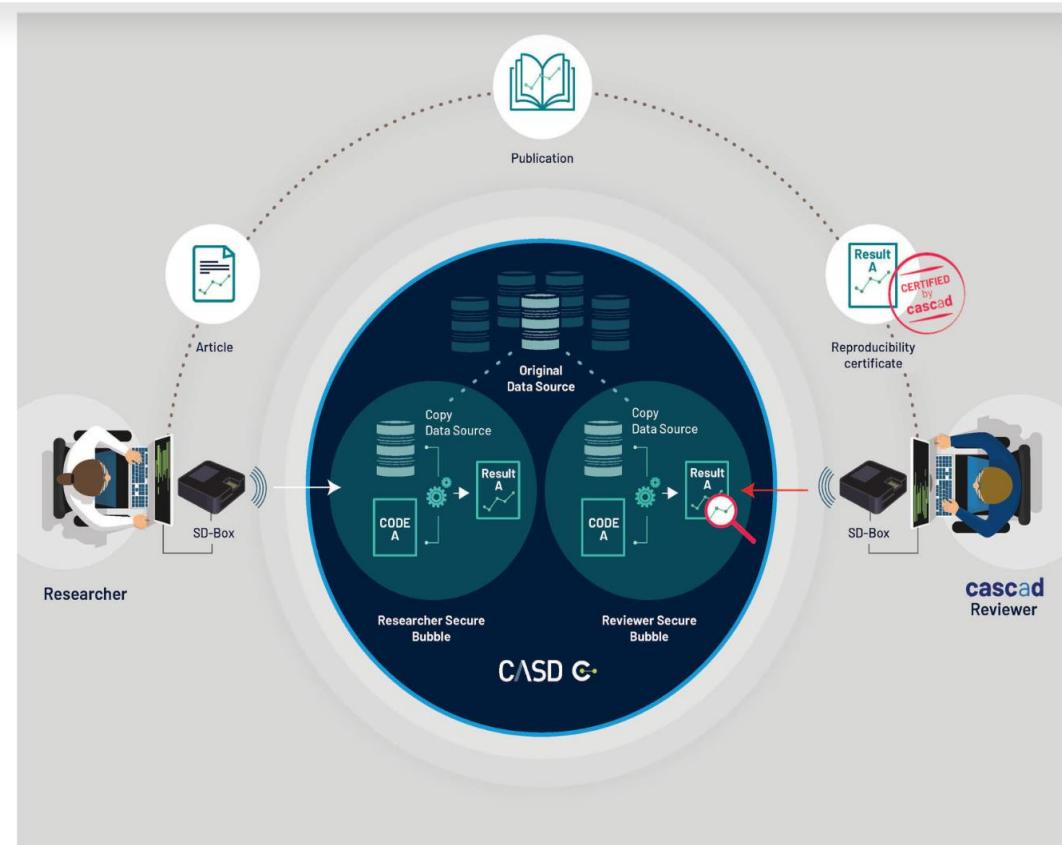




How can others verify reproducibility?

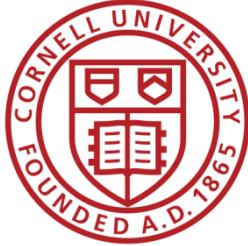


How can others verify reproducibility?



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

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



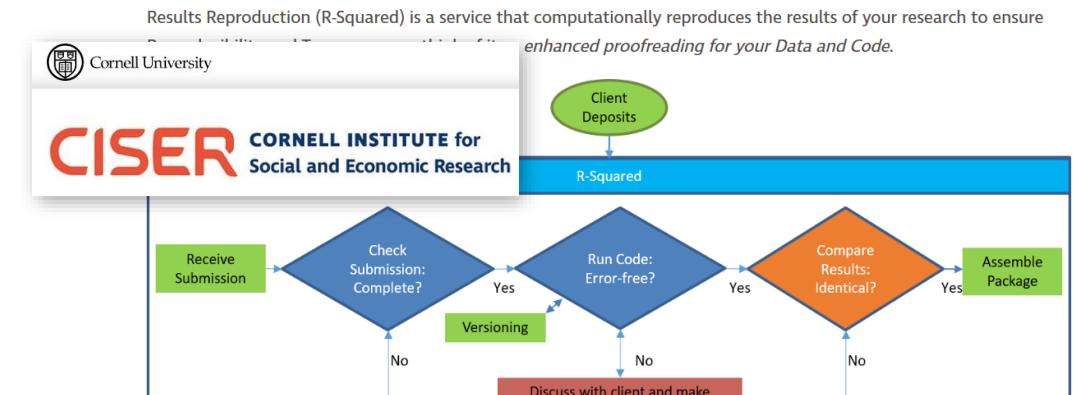
Verification service

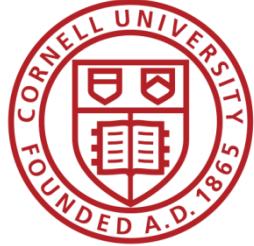
- Permanently accredited staff with access to confidential data
- On-demand, arms-length verification of code and results
- Effectiveness rests on credibility of the service
- Similar to existing services that perform pre-submission verification within universities and research institutes, but with access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



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

RESULTS REPRODUCTION (R-SQUARED)





Embed within disclosure avoidance system



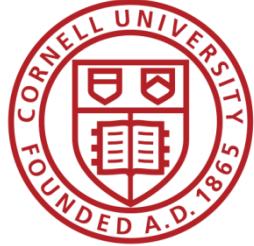
FORSCHUNGSDATENZENTRUM
der Bundesagentur für Arbeit im Institut für
Arbeitsmarkt- und Berufsforschung

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

[Wir über uns](#)

Das Forschungsdatenzen

- 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!**



IAB Certificate?



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



*Certificate of computational
reproducibility*

The following programs were executed:

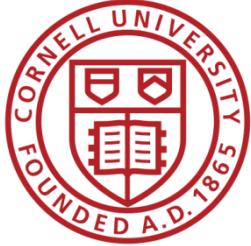
- master.do
- Reg1.do
- Reg2.do
- Graph1.do

and produced the following output:

- Table 1.xlsx
- Table 2.csv
- Figure 1-data.csv

Checksum of all files:

- SHA256: 1234567890ABCDE
 - PK Signature: 12345 ABCDE 56789 BCDE1
- Our public key can be found on our website at doi.org/10.5164/IAB.PK.2019.v1.



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

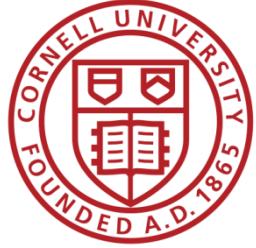
Guidance for authors wishing to create data and code supplements, and for replicators.

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

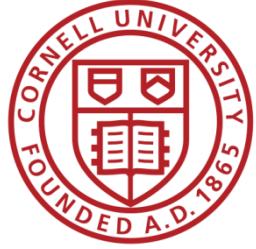


Other options

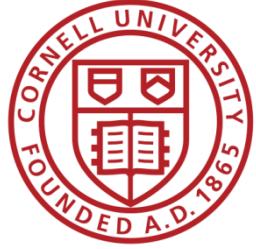
- Cryptographic signatures
- Blockchain
- All of the previous



Image by [VIN JD](#) on Pixabay



This is an unsolved problem.

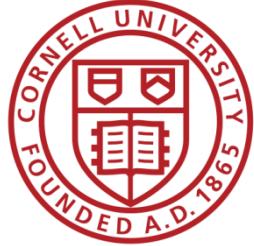


Challenges

- 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](#)

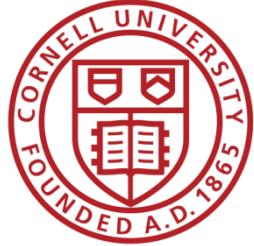
[Author information ▶](#) [Copyright and License information ▶](#)

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 ▶](#)

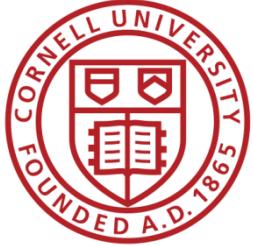
Abstract

[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 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 demographic and industry detail

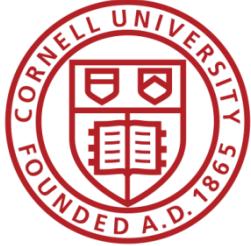
Metrics 4 Downloads

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

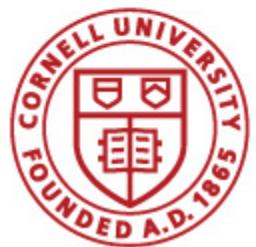
If you use these data, please add this citation to your scholarly resources. Learn about [Data Citation Standards](#).

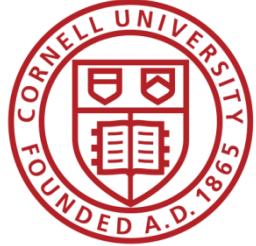
Description	Content
	The Quarterly Workforce Indicators are local labor market data produced and released every quarter by the U.S. Bureau of Labor Statistics. Unlike any other local labor market series produced in the U.S. or the rest of the world, these indicators provide detailed information on job flows for workers (accessions and separations), jobs (creations and destructions) and earnings (by age, race, ethnicity, gender, education, and sex), economic industry (NAICS industry groups), and detailed geography (county, CWA, and metropolitan statistical area). Job flows are estimated from the Longitudinal Employer-Household Dynamics (LEHD) program, which uses administrative records to track the movement of workers between jobs. The indicators also include experimental, unreleased block-level estimates. Job flows are used to construct the first national series of quarterly estimates of gross employment and job flows. These are important enhancements to existing series because they include demographic and industry detail, and are compiled from data that have been integrated at the micro-level by the Longitudinal Employment Dynamics (LED) program.



We went back, archived it, linked it back

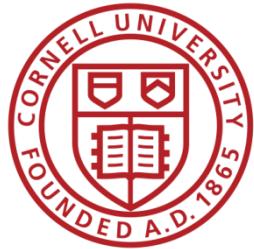
	Dataverse	About
Keyword	Employment Dynamics	
Topic Classification	Economics	
Related Publication	John M. Abowd and Lars Vilhuber, "National estimates of gross employment and job flows from the Quarterly Workforce Indicators with demographic and industry detail," Journal of Econometrics, vol. 161, iss. 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 Workforce Indicators 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 Workforce Indicators with demographic and industry detail (with color graphs)," Center for Economic Studies, U.S. Census Bureau, Washington, D.C., 2011. 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



[README.txt](#)

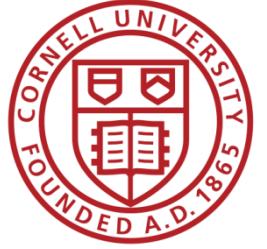
File Type

text/x-web-m

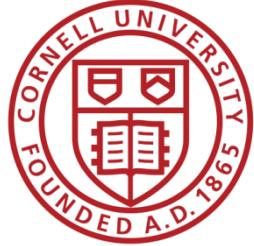
Project Citation:

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?

- Findable,
- Accessible,
- Interoperable, and
- Re-usable

The FORCE11 logo features a blue circular icon with a white target-like pattern next to the word "FORCE11". Below it is the tagline "The Future of Research Communications and e-Scholarship". A navigation bar below the logo includes "ABOUT", "COMMUNITY", and "CODE OF CON".

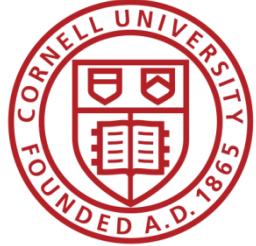
FORCE11 » Groups » The FAIR Data Principles

THE FAIR DATA PRINCIPLES

JOIN IN THE DISCUSSION - LEARN
FAIR Data Principles

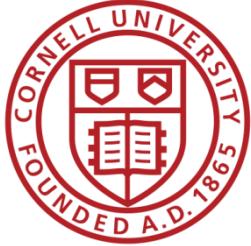
Preamble

One of the grand challenges of data-intensiv



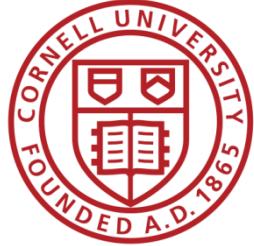
Preserving data in restricted access centers

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



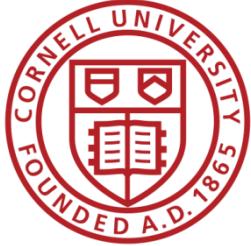
Use existing procedures

- 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



RDCs: Create infrastructure to support

- Publish/ archive/ DOI for code/data releases?
- Publish (version!) preservation policy for
 - Accessible files (“BHP 2016 v1” – how long when v2 is released?)
 - Researcher-generated files (How long are project files kept after end of project)
- Record such policies (re3data.org)
- Provide DOI for archives of researcher-generated files (-> public landing pages -> FAIR)



Some guidance

UNF:6:Upe25NYAZwR+6VsDt5X2lQ==

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)

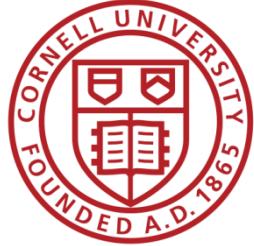
A few guidelines

Social Science Data Editors

Data and Code Guidance by Data Editors

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

- [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_information_hosting.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](#)

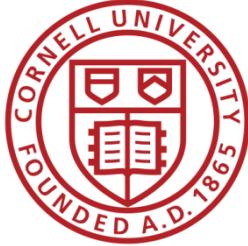
[Centers, Institutes, Programs](#)

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

IDEAS Economic literature Authors Institutions Rank

Twitter Facebook Google+ LinkedIn YouTube Email Print

Center for Economic Studies, U.S. Census Bureau

CES Technical Notes Series

Search within this serial

Publisher Info Serial Info Content Corrections

Content

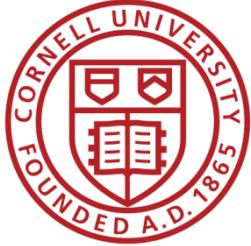
2019

19-02 Augmenting the LBD with Firm-Level Revenue
by John Haltiwanger & Ron Jarmin & Robert Kulick & Javier Miranda & Veronika Pencica

19-01 Duplicate Records and Cross-Year Identifier Inconsistency in Census Year Si
by Lucas Threinen

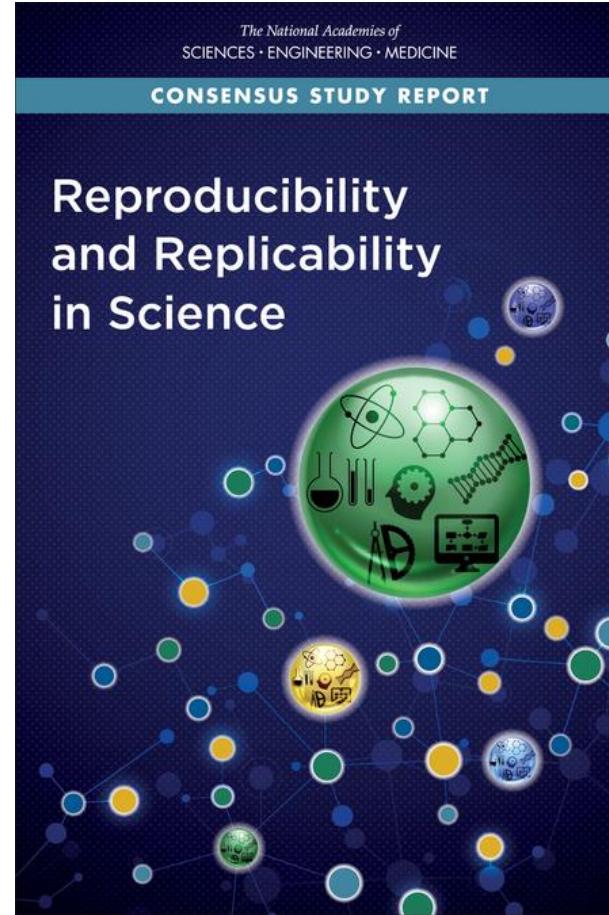
2018

18-03 The Creation and Use of the SIPP Synthetic Beta v7.0

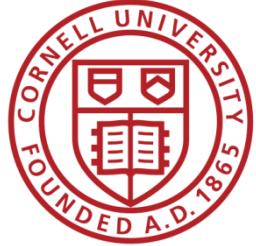


Even better: repositories

- 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.)

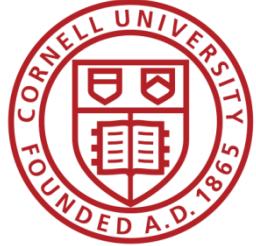


<https://doi.org/10.17226/25303>



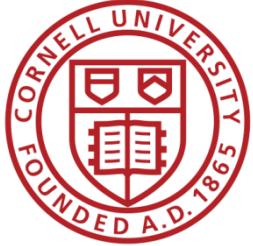
Evolving Journal and Data Infrastructure

Treat all archives
symmetrically!



Evolving Journal and Data Infrastructure

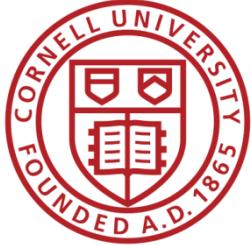
Goal: Use any
repository!
(subject to conditions)



Data (and Code) Availability Statements

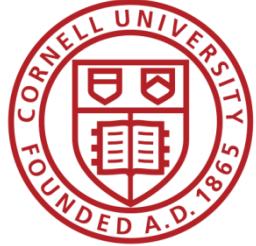
- A statement about where data supporting the results reported in a published article can be found
 - including unique identifiers linking to publicly archived datasets analyzed or generated during the study.
- DASs can increase transparency by providing a reason why data cannot be made (immediately)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

Some
positives



Some random notes

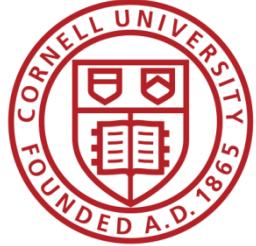
- **Pre-release verification** conducted within the disclosure avoidance mechanism?
- Analogy between **grant** or **RDC proposal** and **pre-registration**
- Incentives of stats agencies: **transparency = credibility**
- From **pre-acceptance verification** to
pre-submission verification (university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



Registered Reports

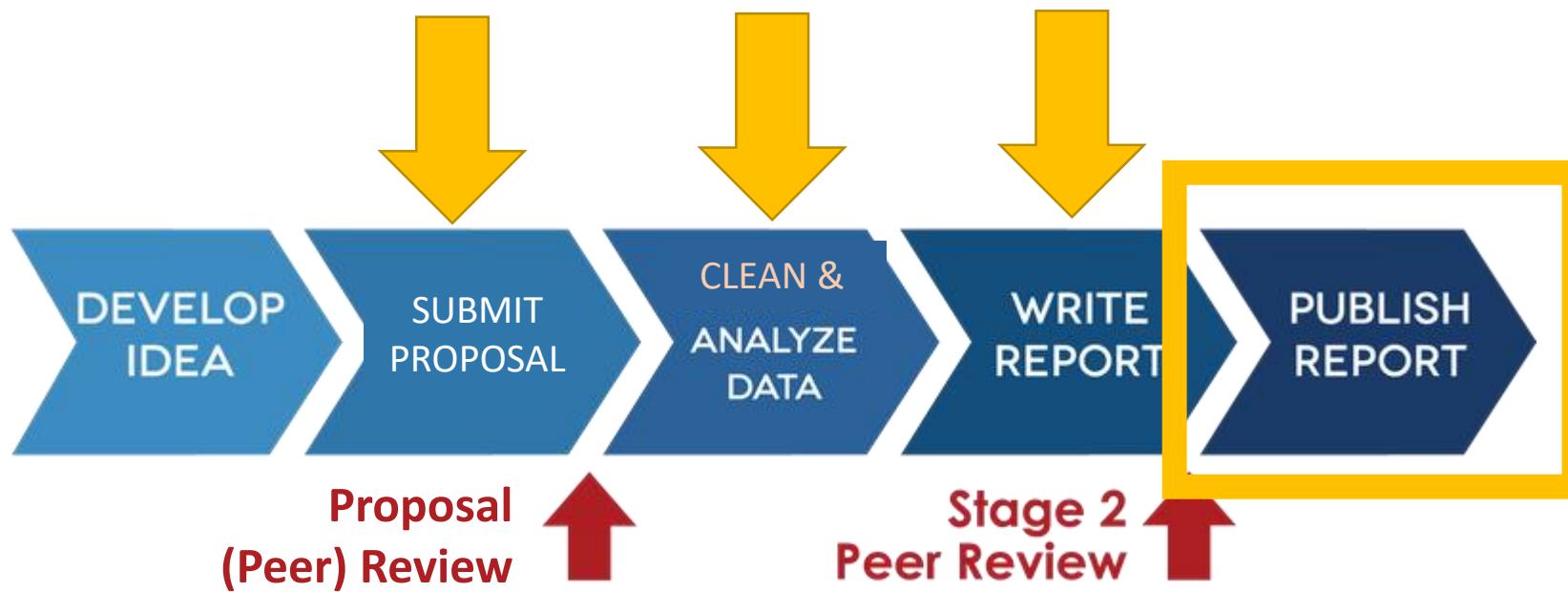
- Close cousin: Results-blind review

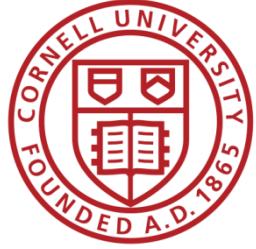




Registered Reports

- Close cousin: **RDC access mechanism**

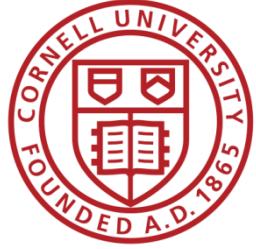




Interesting thought

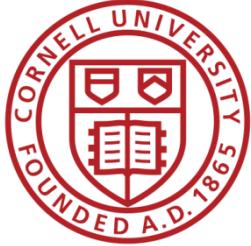
Could the RDC proposal and access mechanism be combined with a registered-report mechanism?

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

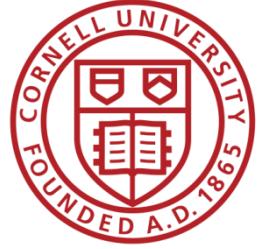
See [Verification guidance](#)

FAQ

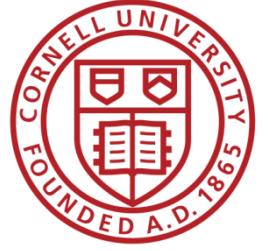
See our growing [FAQ](#). If you have questions or answers to add, please notify us by creating a [new issue](#).

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

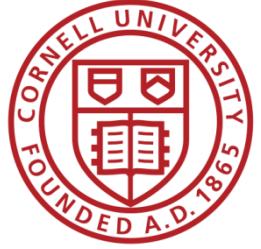
Challenges?



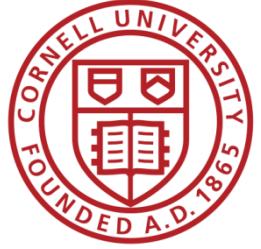
You...



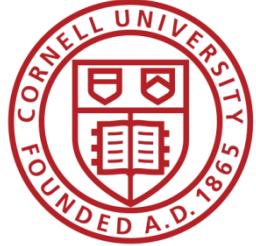
Me...



Change ingrained habits...



New skills to learn...



Shape Your D

Join us at DockerCon 2019,
things Kubernetes, microser

Turn a Git repo into a collection of
notebooks

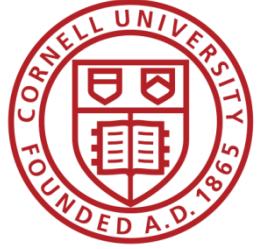
The collage consists of four distinct sections:

- Docker:** A blue banner with the Docker logo and the text "Shape Your D". Below it, a snippet of text reads "Join us at DockerCon 2019, things Kubernetes, microser".
- Why Docker:** A screenshot showing a "Why Docker" section with a "Private Untitled Capsule" interface. It displays a file tree with "environment", "code", and "data Manage Datasets". Below the tree are "Upload" and "Start with Sample Files" buttons.
- R Markdown:** A screenshot of an R Markdown document titled "R Markdown" from R Studio. The document content includes "from R Studio", a preview of a plot titled "Diversity gradient", and the Python logo followed by the word "python".
- Python:** A screenshot of a Python code editor. The code shown is:

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

Below the collage, there are two large blue callout boxes:

- Welcome to RStudio**: "Do, share, teach and learn da"
- Turn a Git repo into a collection of notebooks**



New methods to use ...

Search



UQAM > Archipel

Delete

Répertoires Facultés Bibliothèques

Université de Montréal

New upload

Instructions: (i) Upload minimum one file or fill-in the form below.

Files

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 access the URL on this server."
```

Accueil

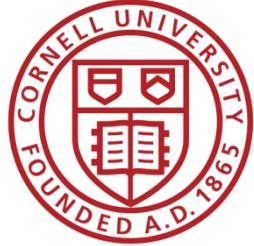
Recherche

Add a file

MON COMPTE

Ouvrir une session

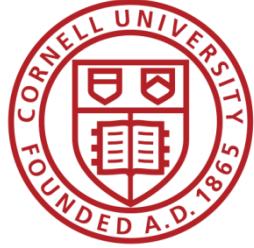
Nouvel utilisateur?



Researchers: New skills to learn/teach

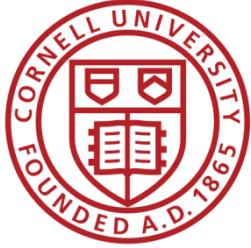
- How to **incorporate reproducible practices** into your workflow
- When to **pre-register**, and when not to
- **Document** early, and often (better READMEs!)
- How, where, and when to **archive data and code**
- How to **license** your contributions!

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)